

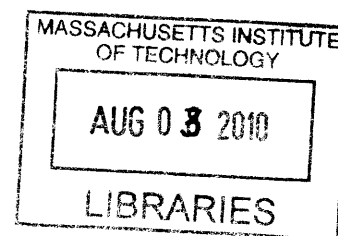
ARCHIVES

# Patterns of Heart Attacks

by

Kimberly N. Shenk

B.S. Mathematics and Operations Research  
United States Air Force Academy, 2008



SUBMITTED TO THE SLOAN SCHOOL OF MANAGEMENT IN  
PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN OPERATIONS RESEARCH  
AT THE  
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2010

Copyright ©2010 Kimberly N. Shenk. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic  
copies of this thesis document in whole or in part.

Signature of Author: \_\_\_\_\_  
Sloan School of Management  
Interdepartmental Program in Operations Research  
May 14th, 2010

Certified by: \_\_\_\_\_  
Dr. Natasha Markuzon  
The Charles Stark Draper Laboratory, Inc.  
Technical Supervisor

Certified by: \_\_\_\_\_  
Professor Dimitris J. Bertsimas  
Boeing Professor of Operations Research  
Thesis Advisor

Accepted by: \_\_\_\_\_  
Professor Patrick Jaillet  
Edmund K. Turner Professor  
Co-Director, Operations Research Center

Report Documentation Page				Form Approved OMB No. 0704-0188	
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE <b>JUN 2010</b>		2. REPORT TYPE		3. DATES COVERED <b>00-00-2010 to 00-00-2010</b>	
4. TITLE AND SUBTITLE <b>Patterns of Heart Attacks</b>				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) <b>Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA, 02139</b>				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT <b>Approved for public release; distribution unlimited</b>					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT <b>Myocardial infarction is a derivative of heart disease that is a growing concern in the United States today. With heart disease becoming increasingly predominant, it is important to not only take steps toward preventing myocardial infarction, but also towards predicting future myocardial infarctions. If we can predict that the dynamic pattern of an individual's diagnostic history matches a pattern already identified as high-risk for myocardial infarction, more rigorous preventative measures can be taken to alter that individual's trajectory of health so that it leads to a better outcome. In this paper we utilize classification and clustering data mining methods concurrently to determine whether a patient is at risk for a future myocardial infarction. Specifically, we apply the algorithms to medical claims data from more than 47,000 members over five years to: 1) find different groups of members that have interesting temporal diagnostic patterns leading to myocardial infarction and 2) provide out-of-sample predictions of myocardial infarction for these groups. Using clustering methods in conjunction with classification algorithms yields improved predictions of myocardial infarction over using classification alone. In addition to improved prediction accuracy, we found that the clustering methods also effectively split the members into groups with different and meaningful temporal diagnostic patterns leading up to myocardial infarction. The patterns found can be a useful profile reference for identifying patients at high-risk for myocardial infarction in the future.</b>					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT <b>Same as Report (SAR)</b>	18. NUMBER OF PAGES <b>76</b>	19a. NAME OF RESPONSIBLE PERSON
a. REPORT <b>unclassified</b>	b. ABSTRACT <b>unclassified</b>	c. THIS PAGE <b>unclassified</b>			



**[This Page Intentionally Left Blank]**



# **Patterns of Heart Attacks**

by

Kimberly N. Shenk

Submitted to the Sloan School of Management on  
May 14th, 2010 in partial fulfillment of the requirements for the  
Degree of Master of Science in Operations Research

## **Abstract**

Myocardial infarction is a derivative of heart disease that is a growing concern in the United States today. With heart disease becoming increasingly predominant, it is important to not only take steps toward preventing myocardial infarction, but also towards predicting future myocardial infarctions. If we can predict that the dynamic pattern of an individual's diagnostic history matches a pattern already identified as high-risk for myocardial infarction, more rigorous preventative measures can be taken to alter that individual's trajectory of health so that it leads to a better outcome. In this paper we utilize classification and clustering data mining methods concurrently to determine whether a patient is at risk for a future myocardial infarction. Specifically, we apply the algorithms to medical claims data from more than 47,000 members over five years to: 1) find different groups of members that have interesting temporal diagnostic patterns leading to myocardial infarction and 2) provide out-of-sample predictions of myocardial infarction for these groups. Using clustering methods in conjunction with classification algorithms yields improved predictions of myocardial infarction over using classification alone. In addition to improved prediction accuracy, we found that the clustering methods also effectively split the members into groups with different and meaningful temporal diagnostic patterns leading up to myocardial infarction. The patterns found can be a useful profile reference for identifying patients at high-risk for myocardial infarction in the future.

Technical Supervisor: Dr. Natasha Markuzon  
Title: Principal Member of the Technical Staff  
The Charles Stark Draper Laboratory, Inc.

Thesis Advisor: Professor Dimitris J. Bertsimas  
Title: Boeing Professor of Operations Research  
Co-Director, Operations Research Center  
Massachusetts Institute of Technology

**[This Page Intentionally Left Blank]**

# Acknowledgements

Throughout my journey here at MIT, many people have been there to encourage, help and advise me. Therefore, I would like to take the time to thank these people for their positive impact on my life.

First and foremost, I thank God for the amazing path he has and is sending me down. He has blessed me with the opportunity to attend MIT and I am excited to see his plans for how I will use this education in the future. I know that his perfect plan for us is better than anything we try to devise on our own and greater than anything we can imagine. Proverbs 3:5-6.

Second, I would like to thank my two advisors: Dr. Natasha Markuzon of Draper Laboratory and Professor Bertsimas of the Operations Research Center at MIT. I am very grateful for all of the time and thought you put into this research. You both provided invaluable advice and insight and without them, the completion of this thesis would have been impossible. I appreciate your patience through the trying times and your genuine belief that I could succeed.

Furthermore, I would like to thank Sanjay Ghimire and Dr. Nathan Gunn from Verisk Health. I appreciate the time they took out of their busy schedules to provide extensive data support for this research.

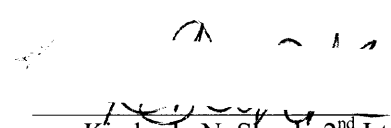
To my friends and study group: Tommy, Chris, Scott, and Ashley. Without your help and contributions on all of those problem sets and projects, I would never have made it through the academics here!

Finally, I'd like to thank my family for their unconditional love and support. To my husband Peter, I know it was difficult being apart for the past two years, but the struggles have only made us stronger. Thank you for being my backbone and my rock; without you I would not have been strong enough to make it through MIT. God has amazing plans for us and I cannot wait to start our next journey, together! To my mom, dad, and Ryan: thank you for all of the advice and prayers before I came here and while I was here. I am so blessed to come from such a supportive, accepting and loving family. Ryan, you mean the world to me; thanks for always being my best friend.

This thesis was prepared at The Charles Stark Draper Laboratory, Inc., under Internal Company Research Project 22928-001 and 223985-001, Data Mining for Draper.

Publication of this thesis does not constitute approval by Draper or the sponsoring agency of the findings or conclusions contained herein. It is published for the exchange and stimulation of ideas.

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or The U.S. Government.

  
Kimberly N. Shenk, 2<sup>nd</sup> Lt., USAF

4 May 2010  
May 14, 2010

**[This Page Intentionally Left Blank]**

# Contents

<b>1 Introduction .....</b>	<b>12</b>
<b>2 The Data .....</b>	<b>16</b>
2.1 Data Overview .....	17
2.2 Data Aggregation .....	18
2.3 Shifting Time-Series Data.....	19
2.4 Cost Bucket Partitioning.....	21
2.5 Training and Test Set Selection .....	22
<b>3 Methods .....</b>	<b>24</b>
3.1 Supervised Learning: Random Forest.....	24
3.1.1 Performance Metrics .....	24
3.1.2 Random Forest Algorithm.....	25
3.2 Unsupervised Learning: Clustering Methods .....	26
3.2.1 K-means Clustering .....	27
3.2.2 Spectral Clustering .....	27
3.3 Baseline Performance .....	27
3.4 Joint Methodology .....	28
<b>4 Prediction Results .....</b>	<b>31</b>
4.1 Baseline Performance .....	31
4.2 Performance Using Clustering .....	32
<b>5 Patterns of Interest.....</b>	<b>36</b>
5.1 Pattern of gradually increasing occurrence of COPD .....	37
5.2 Pattern of Deteriorating Health.....	42
5.3 Pattern of significant occurrence of Chest Pain 3 Months before MI.....	45
5.4 Pattern of significant occurrence of CAD 3 Months before MI .....	50
5.5 Pattern of significant occurrence of Chest Pain 6 Months before MI.....	52
5.6 Pattern of gradually increasing occurrence of Anemia.....	54
5.7 Pattern of gradually increasing occurrence of Cerebrovascular Disease.....	56
5.8 Patterns associated with Diabetes, Hypertension, and Hyperlipidemia.....	58
5.9 Patterns associated with No MI .....	60

5.10 Summary of Results .....	63
<b>6 Conclusion and Future Research.....</b>	<b>65</b>
<b>References .....</b>	<b>67</b>
<b>Appendix A.....</b>	<b>71</b>
<b>Appendix B.....</b>	<b>73</b>
<b>Appendix C.....</b>	<b>75</b>
<b>Appendix D.....</b>	<b>77</b>

# List of Tables

Table 2.1 Summary of Aggregated Variables.....	19
Table 2.2 Summary of the variables used in this study .....	21
Table 2.3 Cost bucket partition summary .....	22
Table 4.1 Random Forest Classification Results .....	31
Table 4.2 Cost Bucket 1 Spectral Clustering Results .....	33
Table 4.3 Cost Bucket 2 Spectral Clustering Results .....	33
Table 4.4 Cost Bucket 2 K-means Clustering Results .....	34
Table 4.5 Cost Bucket 3 Spectral Clustering Results .....	34
Table 5.1 Cost Bucket 2 Spectral Cluster 7 Diagnoses .....	39
Table 5.2 Cost Bucket 2 K-means Cluster 8 Diagnoses .....	42
Table 5.3 Cost Bucket 2 K-means Cluster 1 Diagnoses .....	45
Table 5.4 Cost Bucket 2 Spectral Cluster 1 Diagnoses .....	47
Table 5.5 Cost Bucket 2 K-means Cluster 7 Diagnoses .....	49
Table 5.6 Cost Bucket 2 Spectral Cluster 6 Diagnoses .....	52
Table 5.7 Cost Bucket 2 Spectral Cluster 8 Diagnoses .....	54
Table 5.8 Cost Bucket 3 Spectral Cluster 4 Diagnoses .....	56
Table 5.9 Cost Bucket 3 Spectral Cluster 5 Diagnoses .....	58

# List of Figures

Figure 2.1 Visual Timeline Representation .....	20
Figure 3.1 Example of Classification Tree Structure.....	26
Figure 3.2 Visual Representation of Joint Methodology .....	29
Figure 5.1 Bucket 2 Spectral Cluster 7 Temporal Pattern .....	38
Figure 5.2 Bucket 2 K-means Cluster 8 Temporal Pattern .....	41
Figure 5.3 Bucket 2 K-means Cluster 1 Temporal Pattern .....	43
Figure 5.4 Bucket 2 Spectral Cluster 1 Temporal Pattern .....	46
Figure 5.5 Bucket 2 K-means Cluster 7 Temporal Pattern .....	48
Figure 5.6 Bucket 2 Spectral Cluster 6 Temporal Pattern .....	51
Figure 5.7 Bucket 2 Spectral Cluster 8 Temporal Pattern .....	53
Figure 5.8 Bucket 3 Spectral Cluster 4 Temporal Pattern .....	55
Figure 5.9 Bucket 3 Spectral Cluster 5 Temporal Pattern .....	57
Figure 5.10 Bucket 3 Spectral Cluster 10 Temporal Pattern .....	59
Figure 5.11 Bucket 2 Spectral Cluster 3 Temporal Pattern .....	61
Figure 5.12 Bucket 2 Spectral Cluster 9 Temporal Pattern .....	62



# Chapter 1

## Introduction

Myocardial infarction (also known as MI or a heart attack) is a derivative of heart disease that is a growing concern in the United States today. The findings by the Centers for Disease Control and Prevention claim that the leading cause of death in 2006 in the United States was heart disease, with 631,636 deaths [1]. According to the Heart Disease and Stroke Statistics – 2010 Update from the American Heart Association, the estimated annual incidence of MI is 610,000 new attacks and 325,000 recurrent attacks [2]. This means that almost every 34 seconds an American will suffer from a heart attack. With heart disease becoming increasingly predominant, it is important to not only take steps towards preventing MI, but also towards predicting future MI. If we can predict that the dynamic pattern of an individual's diagnostic history matches a pattern already identified as high-risk for MI, more rigorous preventative measures can be taken to alter that individual's trajectory of health so that it leads to a better outcome.

MI is the death of the heart muscle due to the sudden blockage of a coronary artery by a blood clot [3]. When the coronary artery is blocked, it deprives the heart muscle of blood and oxygen. If the blood flow to the heart muscle is not restored in time, the heart muscle will begin to die. The most common symptom of a heart attack is angina pectoris (also known as angina or chest pain). Other common symptoms include shortness of breath, heartburn, arm pain, jaw pain, toothache, headache, nausea and/or vomiting, and upper back pain. However, approximately 25 percent of all heart attacks are silent, meaning there are no symptoms [3]. Hyperlipidemia (high

blood cholesterol) is one of the main factors that increase the risk of developing atherosclerosis and MI. Hypertension (high blood pressure), smoking, family history of heart disease, and diabetes mellitus (both types 1 and 2) are also factors that increase the risk of MI. These symptoms and risk factors will be used later to assemble the data in this study.

MI is an adverse event that is hard to predict, prevent, and even diagnose. Researchers are still trying to determine the clinical characteristics of less typical presentations of patients in whom MI was missed, even after admission to the emergency room [4]. The ambiguity behind some MI events presents the need for understanding patterns in patient's diagnostic histories that are linked to MI. One study attempts to attain this understanding using observational data from ultrasound images of the common carotid artery [5]. The study found that increased common carotid intima-media thickness is associated with future cardiovascular events such as myocardial infarction. Although this procedural assessment is noninvasive and has shown to successfully predict a heart attack, the underlying problem still remains: how to determine who should have this procedure done. It is infeasible to perform this test on everyone and a threshold is still needed to determine who is at risk. One way to find this is through the analysis of temporal diagnostic data from administrative databases.

The amount of health insurance claims data collected, recorded, and filed into databases is vast. Large databases containing medical information on millions of people are very scarce and therefore insurance claims databases are more frequently being used to evaluate clinical outcomes, adverse events, and future health care expenditures [6, 7, 8, 9]. A few of the many advantages of insurance claims data include: the sample size of patients are very large and are geographically diverse, the data contains multiple records on patients over extended periods of time, the data is already collected and available at a low cost, the target population can easily be defined, and there is an absence of reporting bias [10, 7]. However, many have questioned the suitability of insurance claims information for use in medical research because the data is not collected specifically for clinical care analysis but rather for financial reimbursement purposes. There have been findings that suggest insurance claims data fail to identify prognostically important conditions because the data lacks important details on diagnostic and prognostic information that are commonly captured in the traditional clinical information system [10]. Insurance claims data do not report the outcomes from medical care received or changes in the patient's functional status, sometimes making it difficult to use insurance claims data to effectively measure the outcomes of care. Other studies claim that insurance claims contain variability because of the lack of physician uniformity in naming patient conditions [11]. Nevertheless, many studies adequately combat these limitations and successfully show how claims data can be used in medical research [7, 12, 9, 6].

One study by Wennberg et al. on the outcome of prostatectomy compensates for the lack of information regarding outcomes of medical care by utilizing active participation of practicing physicians to help improve the database [7]. The study combined the insurance claims from the Medicare and the Manitoba Health Services Commission claims database with hospitalization

records and dates of death from the Medicare enrollment files or registry files. With this more robust database, the incidence of mortality and nonfatal outcomes following prostatectomy surgery were successfully measured. In another study done on the Veterans Health Administration's discharge database, medical record abstraction data was compared to the accuracy of the ICD-9-CM codes in the Patient Treatment File [12]. The study validated the use of health insurance claims by finding that the probability a MI was present on admission given that it was coded in the Patient Treatment File was high. Other examples include the use of ICD-9-CM codes to detect adverse events such as adverse drug events, surgical adverse events, misadventures, infections, device events, and other adverse events [9], and measuring outcomes of medication adherence based on medical claims [6].

There are many different trajectories of health that can lead to a heart attack. In this paper we utilize data mining methods to present and discuss ways to determine whether a patient is at risk for a future heart attack based on the patient's past history of diagnoses. Specifically, we employ clustering and classification algorithms concurrently on medical claims data from more than 47,000 members over five years to: 1) provide improved out-of-sample predictions of MI, 2) find groups of members that have new and interesting temporal diagnostic patterns leading to MI and 3) show our method is a systematic way for finding patterns already known to lead to MI and can be applied to other diseases and adverse events. We find that our methods effectively split the members into groups with different clinical characteristics and uncover several interesting dynamic diagnostic patterns. These diagnostic patterns provide meaningful insights into what diagnosis combinations and profiles classify a patient as high-risk for MI and are strong predictors of a future MI.

The rest of this paper is summarized as follows: In Section 2, we describe the data and the criteria we use for its aggregation. In Section 3, we present both the supervised classification algorithm and unsupervised clustering methods we use. In Section 4, we report the performance of the MI predictions. In Section 5, we provide an in-depth discussion of the diagnostic patterns uncovered. Finally, in Section 6 we briefly discuss our conclusions and provide suggestions for future research.

**[This Page Intentionally Left Blank**

# Chapter 2

## The Data

When an individual receives care in a hospital or from a professional healthcare provider, information about the service is recorded in order for the healthcare provider to receive compensation. A medical coder is a medical records and health information technician that specializes in translating and codifying the information regarding the patient's visit to the doctor into a more universal medical record [13]. A set of published codes is used to assign a code to each diagnosis, procedure, and prescription drug that is documented by the healthcare provider. The universal codes then allow the insurance company to determine how much the health care provider will be reimbursed. Therefore, a health insurance claim is a codified bill that the healthcare provider sends to the insurance company to receive payment for their services.

There are two different types of health insurance claims: medical claims and pharmacy claims. The main elements of a medical claim are the diagnosis and procedure codes while a pharmacy claim contains prescription drug codes. There are different coding systems used for each of these coded elements. Diagnostic codes group diseases, disorders, symptoms, and medical signs. Procedure codes identify the specific health actions taken by professional health care providers. The diagnosis and procedure data is coded using the ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) codes [14]. ICD-9-CM is the universal coding system created by the U.S National Center for Health Statistics (NCHS) for assigning codes to diagnoses and procedures. Volumes 1 and 2 are used for diagnostic codes while Volume 3 is a scheme of procedural codes. There are over 17,000 individual diagnostic ICD-9-CM codes and

over 25,000 individual procedure ICD-9-CM codes. We refer the reader to Appendix A for a summary of the ICD-9-CM codes. Pharmacy claims are coded using prescription drug codes from the National Drug Code (NDC) system [15]. NDC drug codes include some over-the-counter drugs, insulin formulations, prescription drugs, and herbal drugs distributed in the United States. There are over 317,000 individual NDC drug codes.

In addition to the medical, procedure, and drug codes, health insurance claims include other important information pertaining to place of service, cost, and patient demographics. There are approximately 100 different codes for specifying the location the care was received. A detailed account of the dollar amount the insurance company reimbursed the health care provider for each procedure performed and drug prescribed is documented. The date the diagnosis was given, the procedure was performed, and/or the drug was prescribed is also documented. Finally, information regarding the patient's gender, date of birth, and member identification number are also documented in the health insurance claim.

## **2.1 Data Overview**

The data used in this study is provided by Verisk Health, a company that leverages healthcare data to identify, manage and minimize medical and financial risk at both the patient and population levels. The data is generated from health insurance claims filed for 47,763 members from a commercially insured population across the country over the observation period 01/27/2000 – 11/30/2007. The criteria for member selection for the study population are as follows:

- i. All members must have at least 5 Coronary Artery Disease (CAD) diagnosis codes and 5 hypertension diagnosis codes
- ii. Data intake period for each member must be at least 5 years
  - a. Members should have at least 5 years of continuous eligibility
  - b. The most recent 5 years of data should be taken
- iii. All members must have at least 100 medical claims
- iv. All members must have at least 5 pharmacy claims

These criteria are used to ensure that all members of the study population have the same length of eligibility within the same time period and have continuous coverage. The criteria also help to define the target population and create a denser data set. With these criterion applied, the resulting data set consists of 19,963,685 health insurance records. These health insurance records include 13,217,714 individual medical records and 6,745,971 individual pharmaceutical records for the 47,763 distinct members along with basic demographic information for each member such as member identification number, date of birth, and gender.

## 2.2 Data Aggregation

As discussed earlier, there are over 17,000 diagnosis and 25,000 procedure ICD-9-CM codes and 317,000 drug NDC codes. If every diagnosis, procedure, and drug code were used, the data set would have close to 360,000 attributes. In order to reduce the data set into a more workable size, the diagnosis and procedure codes are broken down into groups of similar codes. The diagnosis codes are reduced into 218 diagnosis groups, the procedure codes are reduced into 180 procedure groups, and the prescription drug codes are reduced into 538 drug groups. The codes for each of these groups were developed by Verisk Health. Refer to Appendix B for an example of how diagnosis codes are reduced into a diagnosis group.

To further simplify the number of attributes, the data used in this study only includes 46 of the 218 diagnosis group codes and no procedure or drug group codes. The 46 diagnosis groups chosen either directly correspond to a myocardial infarction event, can be causes of an MI, or are risk factors for an MI. This simplification was done to eliminate diagnoses unrelated to a Myocardial Infarction; for example, a broken leg. The remaining 172 diagnosis codes are grouped into the variable: Other Diagnosis. See Appendix C for a detailed list of the diagnosis groups used in this study.

Although there are only 47,763 members in the data set, there are over 13 million different medical insurance records. This means that, on average, each patient has approximately 270 different diagnoses recorded over the observation period. To get a better view of each member's past medical history, we want to compress the hundreds of medical records per member into one record per member. This will provide us with a time-series record of the patient's diagnostic profile over the observation period.

To achieve a time-series glimpse of each patient's medical history, the observation period is split into 21 periods, each 90 days in length (see Appendix D). We look at the diagnoses given to the patient during each 90 day period by counting the number of claims filed under each diagnosis. For example, if a patient has 20 different claims filed for the diagnosis of chest pain in period 12, the variable ChestPain\_Period12 would have a value of 20. This allows us to view the diagnostic activity for each patient every 3 months. We also count the number of visits to the hospital that the member has in each period. The other variables include the total medical cost for each period and gender. Table 2.1 below shows a summary of the aggregated variables.

Variable Number	Description
1	Member Identification Number
2	Gender
3 - 49	Diagnosis group counts for period 1
50	Number of Emergency Room Visits for period 1
51	Total Medical Cost for period 1
52- 98	Diagnosis group counts for period 2
99	Number of Emergency Room Visits for period 2
100	Total Medical Cost for period 2
.	.
.	.
.	.
983 - 1029	Diagnosis group counts for period 21
1030	Number of Emergency Room Visits for period 21
1031	Total Medical Cost for period 21

Table 2.1: Summary of Aggregated Variables

### 2.3 Shifting Time-Series Data

The target outcome we want to predict is the occurrence of a myocardial infarction diagnosis and a trip to the emergency room. Our target variable consists of an MI diagnosis code and an ER place of service code because the documentation of a MI diagnosis group code by itself does not necessarily mean a heart attack occurred. The MI diagnosis group code not only accounts for MI events but also for follow-up diagnoses such as post-myocardial infarction syndrome. Therefore, considering an MI diagnosis along with a trip to the ER as our target variable will help to ensure that the target outcome is in fact a heart attack. The target variable is binary: denoted  $\{+1, -1\}$  for the occurrence of an MI and ER event and no occurrence of an MI and ER event, respectively. Since it is possible for patients in the dataset to have more than one episode of MI, we only consider the first occurrence. We found that people experiencing multiple MI events develop co-morbidities that bias their diagnostic history when compared those who have had no previous MI.

For each member with MI and ER that is selected, we randomly select a member with no MI and ER. Therefore, the dataset is balanced with an equal number of heart attack members and non-heart attack members in the target variable. The non-MI member is chosen only if the member's entire diagnostic history up to that point shows no MI event.

After the "yes MI" and "no MI" members are selected, we want to look at the 3 periods (9 months) of diagnostic history leading up to the MI/no-MI event for every member and compare them. However, all of the MI/no-MI events occur across 21 different time periods. To compare



the diagnostic histories of all the patients concurrently, we align the data to make the observations date-independent, thus preserving only the order of events. Figure 2.1 below provides a visual timeline representation of the time-series data shift.

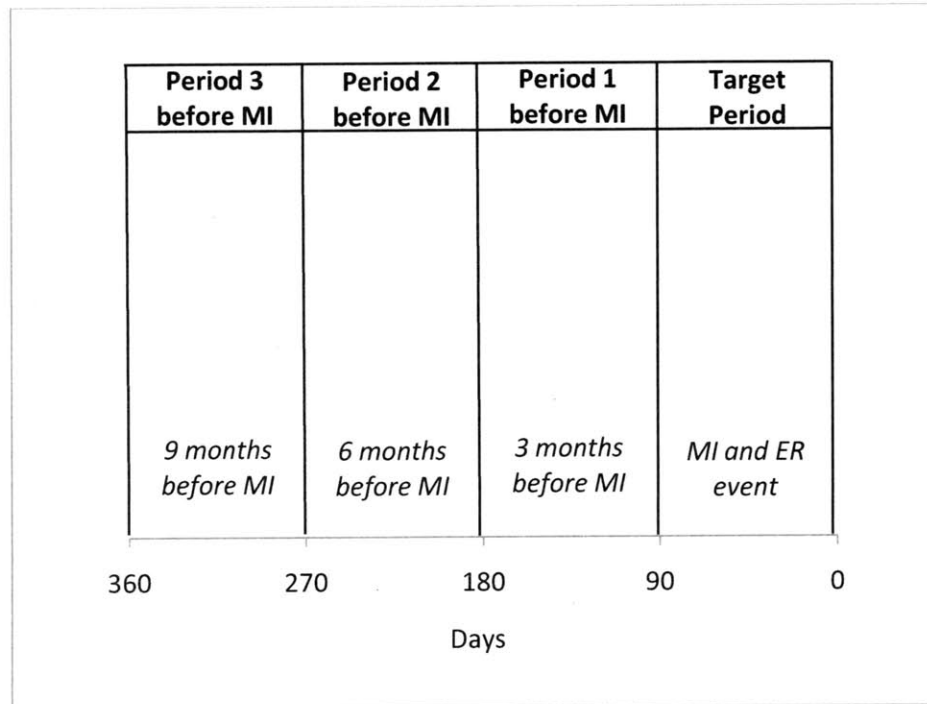


Figure 2.1: All MI events and no-MI events occur in a 90-day target period. The patient's diagnostic history is recorded at 3 months (90-180 days), 6 months (180-270 days), and 9 months (270-360 days) before the MI target period.

The time series shift has simplified the dataset even more, and we now only consider 48 variables per time interval. A summary of these variables is given in Table 2.2 below.

Variable Number	Description
1	Member Identification Number
2	Gender
3 - 49	Diagnosis group counts 9 months before M.I.
50	Total Cost 9 months before M.I.
51 - 97	Diagnosis group counts 6 months before M.I.
98	Total Cost 6 months before M.I.
99 - 145	Diagnosis group counts 3 months before M.I.
146	Total Cost 3 months before M.I.
147	MI and ER target outcome

Table 2.2: Summary of the variables used in this study

## 2.4 Cost Bucket Partitioning

The total cost of medical care in the three 90 day periods leading up to the MI and ER event widely range from \$0 to \$636,508. Examining the cost structure of our data more closely, we find that approximately 70% of the overall cost is generated by only 11% of the population. This means that the high-risk patients with high medical expenses are a very small proportion of the data and could skew our final results. According to the American Medical Association (AMA), only 10% of individuals have projected medical expenses of approximately \$10,000 or greater per year, which is more than four times greater than the average projected medical expense of \$2,400 per year [16]. To lessen the effects of these high-cost outliers we divide the data into three different cost buckets based on the findings by the AMA. Since cost is a good summary of a person's health [29], cost bucket partitioning allows us to perform analysis on patients that have similar conditions of health. Cost bucket 1 represents below average to average health-risk members, cost bucket 2 represents average to above average health-risk members, and cost bucket 3 represents high health-risk members. Table 2.3 below gives a summary of the cost bucket partitions.

BUCKETS	1	2	3
RANGE	< \$2,000	\$2,000 - \$10,000	>= \$10,000
Percentage of the Data	67.56%	21.56%	10.88%
Number of Members Considered	4416	1409	711
Percentage with MI and ER event	36.14%	43.22%	38.12%
Percentage without MI and ER event	63.86%	56.78%	61.88%

Table 2.3: Cost bucket partition summary

Even though a higher percentage of the members fall into cost buckets 1 and 2, the data in these cost buckets is sparse when compared to the data for members in cost bucket 3. This is because members in the first two cost buckets have a less dense history of diagnoses.

## 2.5 Training and Test Set Selection

In order to effectively evaluate our model’s ability to predict myocardial infarction, we ran out-of-sample tests using independent test data. We randomly partition the full dataset into three separate parts. The first part, the training set, is used to develop the model and fit it to the data. The second part, the validation set, is used to adjust the model’s parameters and assess the performance of the prediction model. The last part, the test set, is used to evaluate how the model performs on data that the chosen model has not previously seen – the generalization error [17]. However, we do not have enough data in each cost bucket to set aside a validation set. To handle this problem, we randomly split the data into only a training set and a test set. We use 10-fold cross validation on the training set to obtain 10 different validation trials. The training set is split into 10 equal parts so that for each trial, we use nine-tenths of the training set to train the different models and the remaining one-tenth to calibrate the models and select the best parameters. The model performance is then evaluated using the unseen test set. We ensure the training set is balanced (contains an equal number of  $\{+1, -1\}$  values in the target variable) so that neither class receives preferential treatment by the learning algorithm because of its prevalence in the training set. The test set consists of the data left over and remains unbalanced so that it represents the original distribution of the target variable values in the sample population.

**[This Page Intentionally Left Blank]**

# Chapter 3

## Methods

There are two different forms of statistical learning: supervised learning and unsupervised learning. In supervised learning, we know the target outcome value and we build a model to predict the known output using a set of input feature values. In unsupervised learning, we have no measurement of the outcome variable. This form of learning uses the known feature values to find patterns in the data or find relationships between the feature values. In this study, we will use both forms of statistical learning concurrently. We will show that using supervised learning alone yields weak results, but a joint approach increases the predictability of a heart attack. We will first introduce the supervised and unsupervised learning methods we employ and then discuss the steps and motivation behind our joint methodology.

### 3.1 Supervised Learning: Random Forest

#### 3.1.1 Performance Metrics

To assess the performance of a supervised learning algorithm, many different performance metrics are used. However, not all metrics are appropriate for every area of study because different performance metrics measure different tradeoffs in the predictions made by a classifier [18]. Additionally, supervised learning algorithms can yield strong performance with one metric

while performing weakly with others. We use the metric of accuracy which measures the total number of correct predictions the model makes out of the entire tested population.

### **3.1.2 Random Forest Algorithm**

The data in this study is labeled in advance and therefore using supervised learning to predict the known class label (heart attack/ no heart attack) is appropriate. We choose the random forest algorithm because of its attractive property of detecting variable interactions and excellent performance as a learning algorithm. The algorithm estimates the importance of a variable by looking at how much prediction error increases when data for that variable is permuted while all others are left unchanged [19]. When tested on eight different performance metrics against nine other supervised learning algorithms, random forests gave the second best average performance across all of the metrics and different test problems [18].

The random forest algorithm builds a large collection of uncorrelated classification trees. A classification tree is a decision tree that recursively partitions a dataset into smaller and smaller groups that are similar based on the known class label. A single classification tree has a hierarchical structure similar to the example in Figure 3.1 below. We can predict whether a new member from the test set will have a heart attack by taking the member's information and following the tree down its structure. In Figure 3.1, if the member has CAD and chest pain, we predict a heart attack. However if the member has CAD but does not have chest pain, we predict no heart attack. Similarly, if the member does not have CAD, we predict no heart attack.

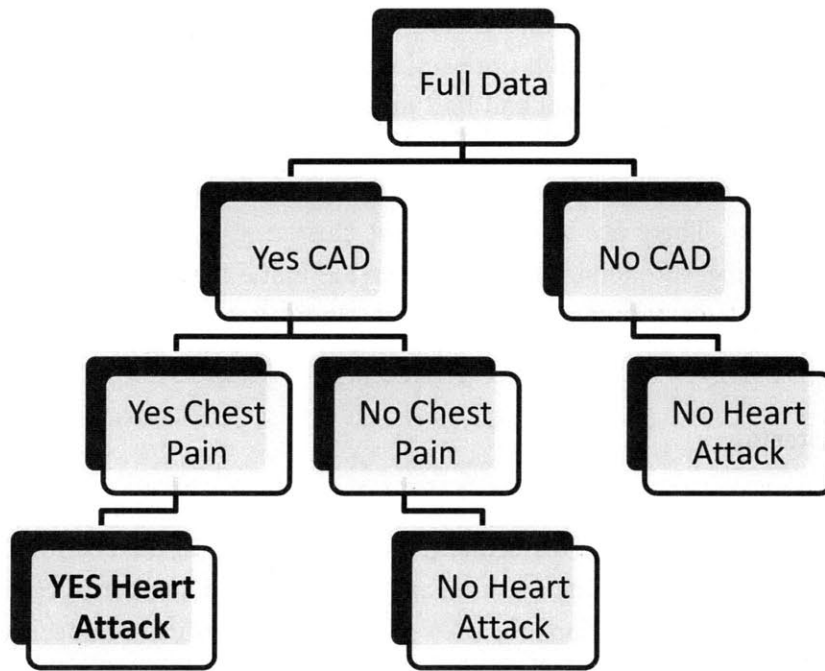


Figure 3.1: Example of a classification tree structure

The many classification trees built by the random forest algorithm are much more complex with many more diagnoses to consider. In the algorithm, a new member from the test set is classified by taking the member's data and following each unique tree down its structure. Each tree will output a classification for that member as "Yes Heart Attack" class or "No Heart Attack" class. The random forest algorithm tallies the number of times the member was classified as either "Yes" or "No" and then classifies the member by choosing the class that has the most votes out of all of the trees in the forest. The parameter that requires tuning in this algorithm is  $t$ , the number of trees to construct. The random forest model is built from the training data using 10-fold cross-validation and different values of  $t$  to find the optimal value of  $t$ . The model with the optimal value of  $t$  is then used on the unseen test set. For more details on the random forest algorithm see [20].

### 3.2 Unsupervised Learning: Clustering Methods

Many times in data analysis we do not know the outcome variable corresponding to a given set of input variables or features. Instead, we are either searching for relationships between the input variables or trying to organize the data into different groups. In this study, we know the outcome variable, heart attack or no heart attack, but we recognize that there are many different trajectories of health that can lead to a heart attack. There is not one set pattern of health or

diagnostic combination that leads a person to a heart attack. Instead, we will show that there are many different dynamic health patterns and time-series diagnostic relationships that can lead to a heart attack. One way to segment the data and find these interesting groups with different health patterns is through cluster analysis. Cluster analysis attempts to divide the data into different groups, called clusters, so that the members of each cluster are more similar to each other than to members of other clusters. There are several different clustering methods, but all of them are fundamentally based on a measure of similarity between the individual members being clustered. We consider two different clustering methods in this study: k-means clustering and spectral clustering.

### **3.2.1 K-means Clustering**

K-means clustering is based on the idea that data can be divided into clusters where each cluster is represented by a center point [21]. This center point, also known as the centroid, is the mean of the data within the cluster. To initiate the k-means algorithm, the number of clusters,  $k$ , must be specified. The initial centroid locations for each of these  $k$  clusters can be specified or chosen at random. Given an initial set of  $k$  centroids, the k-means algorithm assigns data points to the initial clusters by minimizing the squared Euclidean distance from the data point to the centroid of the cluster. Once all points are assigned to the closest centroid, the centroid (the mean) of each cluster is recomputed. The points are then re-assigned to the closest centroid and the centroids are recomputed again. These two steps occur repeatedly until the cluster assignments do not change. In this study we use  $k = 5, 7, 10$  and choose the initial centroid locations randomly. More details on the algorithm can be found in [21].

### **3.2.2 Spectral Clustering**

Many studies have shown that the spectral clustering algorithm is more effective in finding clusters than traditional clustering algorithms such as k-means [22]. Spectral clustering considers the pairwise similarity of data points while k-means only considers the similarity values from the individual data points to their centroids. The algorithm is also attractive because it is very simple to implement and can be solved efficiently by standard linear algebra methods [23]. The algorithm is initiated by measuring the pairwise similarities of the data points and constructing the corresponding similarity matrix. The number of clusters,  $k$ , must also be specified. In this study we use  $k = 5, 7, 10$ . Details on the algorithm can be found in [22, 23]. We will use both k-means and spectral clustering methods to see how our results differ.

## **3.3 Baseline Performance**

Mentioned briefly above, we utilize a joint methodology, with both supervised and unsupervised learning, to predict a heart attack. In order to determine how well our joint method performs, we



devise a baseline method to compare our model’s results to. Our baseline method uses only the random forest supervised learning algorithm to predict a heart attack within each cost bucket. To show that our joint methodology for heart attack prediction is meaningful, we will compare the baseline performance to the performance of executing the random forest supervised learning algorithm and unsupervised learning algorithms in parallel.

### 3.4 Joint Methodology

The most traditional use of supervised learning is in the context of binary classification. This means that given all of the input features describing each member from the sample population, a model is built to classify the members into two separate groups: one group with heart attacks and one group without. However, categorizing the members in this way can be an oversimplification of the problem because it assumes that members falling into the same class have similar health characteristics. Therefore, we apply clustering algorithms to the data to first find different groups of people with different health histories and then try to predict a MI event within each of these sub-populations. Additionally, if interesting groups of people that are more at risk for a heart attack do exist, the algorithms should find these groups. The diagnostic characteristics of patients in these groups can help us to identify high-risk time-series health patterns that lead to heart attacks.

Our method first takes the data from each cost bucket and divides it into training and test sets (as explained in Section 2). The second step uses a clustering algorithm to cluster the training set, with the target values omitted. This gives us a model with a set of  $k$  clusters. Each of the  $k$  clusters is a different group of members that are described by different patterns of health. The data from each of the  $k$  clusters is then used as training data to build  $k$  random forest classification models. These models will be used to predict the occurrence of a heart attack within each group of members. After the classification models are trained and validated using 10-fold cross validation, the left-over test set is then run through the cluster model already built by the training data (target values omitted). The test set, previously unseen by the clustering model, will now also be divided so that each member is assigned to a group that most closely represents that member’s diagnostic history. Now, each cluster consists of members from both the training set and the test set. The test data from each of the  $k$  clusters is then used to test its corresponding classification model. A visual representation of the method is shown in Figure 3.2 below.

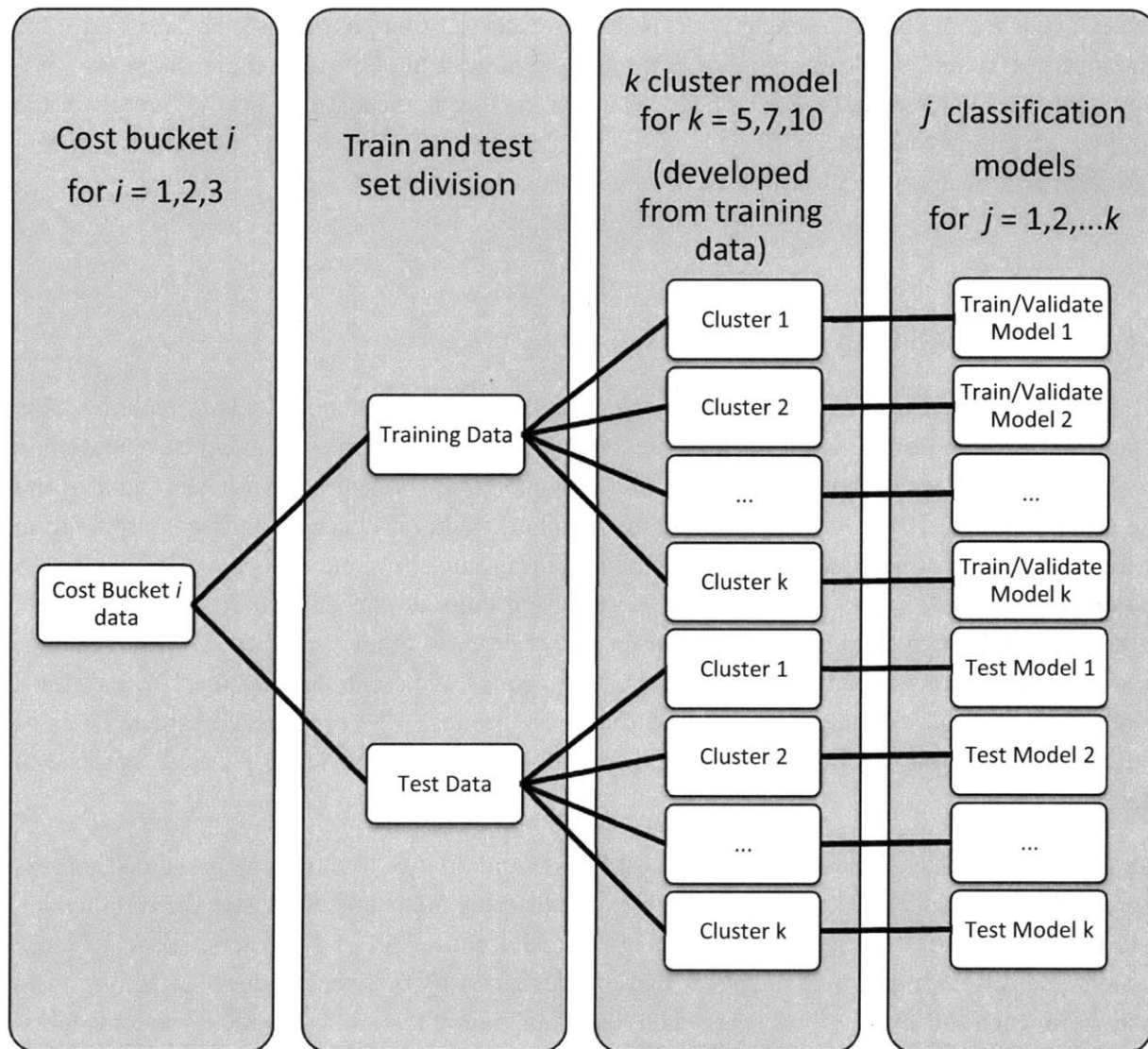


Figure 3.2: Joint method using both supervised and unsupervised learning

The prediction performance of each classification model is then compared to the baseline performance. An increased level of performance from the baseline means that clustering does find interesting groups of people that have certain diagnostic characteristics over the time leading up to a heart attack. Additionally, a strong prediction result for cluster  $k$  indicates that we can successfully predict a heart attack for the types of members that fall in cluster  $k$ .

**[This Page Intentionally Left Blank]**

# Chapter 4

## Prediction Results

### 4.1 Baseline Performance

We first ran the random forest classification algorithm using the randomly selected training set from the data. The model's parameters were tuned using 10-fold cross validation. The model chosen was then used to predict the class of the remaining withheld data in the testing set. We then divided the data into cost buckets and applied the random forest algorithm again. Table 4.1 below shows the prediction results. All prediction results are color coded for viewing convenience. A red tint indicates weak performance, a yellow tint mediocre performance, and a green tint strong performance.

Prediction Rate			
Overall (no cost buckets)	Cost Bucket	Baseline	After Clustering
51.81%	1	49.63%	64.75%
	2	55.99%	72.93%
	3	58.31%	78.25%

Table 4.1: Random forest classification results for each cost bucket.

The first column in Table 4.1 above shows the overall prediction rate before cost bucketing. The third column shows the prediction rates after the data was divided into cost buckets. The random forest model poorly predicts MI in each of the cost buckets but shows a slight improvement from the overall prediction rate. The last column of the table shows the prediction results after clustering was applied to the cost buckets. These prediction results are from the spectral model because it consistently outperformed the k-means model. The prediction rates improve after applying clustering. More details on the clustering method results are discussed in the next section.

## 4.2 Performance Using Clustering

In each cost bucket, the data is divided into a learning sample and a test sample. We first ran both the spectral clustering and k-means clustering algorithms on the learning sample, leaving out the target outcome variable (MI or no MI). The resulting clusters contain members with similar cost and diagnostic characteristics. We then ran the random forest classification algorithm on each cluster separately, using the data in each cluster for training and validation. After the classification models for each cluster were chosen, we took the unseen test sample from each cost bucket (without the target variable) and ran it through the previously developed cluster models. This produced test samples for each cluster. Finally, we assign a prediction to each test set using the previously developed classification models from each cluster. For a more detailed description of this algorithm, refer to Section 3. We ran both the k-means and spectral clustering algorithms for  $k = 5, 7, 10$ . The prediction performance of the random forest algorithm monotonically increased with  $k$ , and therefore we only report the results for  $k = 10$ . Tables 4.2 – 4.5 below outline the cluster structures and prediction performance for each of the resulting clusters by cost bucket. All prediction results are color coded for viewing convenience. A red tint indicates weak performance, a yellow tint mediocre performance, and a green tint strong performance. Rows highlighted in blue indicate clusters that have interesting temporal patterns, relatively high prediction rates, and/or consist of a reasonably large number of members with MI in the target period. Spectral clustering consistently performed better than k-means so we only report the spectral clustering results. However, in cost bucket 2 we show the results from both algorithms for comparison purposes later on.

<b>COST BUCKET 1: Spectral Clustering</b>				
Cluster	YES MI Count	NO MI Count	Total Count	Prediction Rate
1	73	77	150	61.31%
2	42	118	160	83.52%
3	103	81	184	55.00%
4	86	75	161	51.28%
5	52	57	109	57.14%
6	300	245	545	65.40%
7	77	136	213	79.37%
8	91	93	184	59.20%
9	173	148	321	57.53%
10	123	90	213	88.23%
Weighted Average Prediction Rate				64.75%

Table 4.2: The resulting performance per cluster for spectral clustering in cost bucket 1.

<b>COST BUCKET 2: Spectral Clustering</b>				
Cluster	YES MI Count	NO MI Count	Total Count	Prediction Rate
1	32	19	51	72.41%
2	111	116	227	62.58%
3	14	21	35	70.59%
4	21	36	57	82.93%
5	61	50	111	70.97%
6	30	27	57	71.43%
7	30	25	55	68.29%
8	31	14	45	68.00%
9	25	38	63	73.91%
10	45	54	99	70.45%
Weighted Average Prediction Rate				72.93%

Table 4.3: The resulting performance per cluster for spectral clustering in cost bucket 2.



<b>COST BUCKET 2: K-means Clustering</b>				
Cluster	YES MI Count	NO MI Count	Total Count	Prediction Rate
1	22	14	36	85.71%
2	26	29	55	76.92%
3	25	33	58	74.29%
4	4	7	11	-
5	263	272	535	62.33%
6	4	2	6	-
7	20	12	32	85.00%
8	13	8	21	72.22%
9	8	8	16	95.67%
10	15	15	30	81.82%
Weighted Average Prediction Rate				67.17%

Table 4.4: The resulting performance per cluster for k-means clustering in cost bucket 2.

<b>COST BUCKET 3: Spectral Clustering</b>				
Cluster	YES MI Count	NO MI Count	Total Count	Prediction Rate
1	37	50	87	76.12%
2	11	12	23	84.38%
3	10	16	26	87.10%
4	16	8	24	70.83%
5	30	26	56	73.08%
6	12	13	25	75.76%
7	14	14	28	77.78%
8	11	18	29	91.30%
9	14	15	29	86.36%
10	35	18	53	69.57%
Weighted Average Prediction Rate				78.25%

Table 4.5: The resulting performance per cluster for spectral clustering in cost bucket 3.

Tables 4.2 – 4.5 show that the weighted average prediction rate for each cost bucket is considerably greater than the prediction rate found without clustering (Table 4.1) in both clustering algorithms. This shows that clustering does improve our ability to predict MI. Additionally, the weighted average prediction rate dramatically increases from cost bucket 1 to

cost bucket 3 with both clustering algorithms. This shows that overall, we predict MI better for less healthy, more expensive patients. We also find that the spectral clustering algorithm has higher overall weighted average prediction rates per cost bucket than the k-means clustering algorithm.



# Chapter 5

## Patterns of Interest

The clusters resulting from both the spectral and k-means clustering models contain interesting groups of members that have similar temporal diagnostic characteristics. As Tables 4.2 – 4.5 from Section 4 show, some clusters have a large disparity between the number of members who experience MI in the target period and the number of members who do not. Additionally, some clusters have high accuracy for MI prediction while other clusters do not. In order to find interesting temporal patterns of diagnoses that can lead to a heart attack, we analyze the clusters in more depth. The unique and interesting patterns were selected based on three criteria: disparity between the number of members with MI and the number without, accuracy in the prediction of MI, and an interesting diagnostic pattern not usually associated with MI.

In this section, we demonstrate that within each cost bucket there are several different temporal patterns of diagnostic combinations that can lead people to a heart attack. We show that although members from different clusters may have similar cost characteristics and high probabilities of having a heart attack, the diagnostic trajectories of getting to the heart attack are distinct. Many of the trajectories we found have been shown in independent studies to be associated with or lead to MI. For cost bucket 2, we will discuss patterns for patients with Chronic Obstructive Pulmonary Disease (COPD), a pattern for patients with a combination of diagnoses leading to deteriorating health, patterns for patients with significant occurrence of chest pain three months before the MI, patterns for patients with significant occurrence of Coronary Artery Disease (CAD) three months before the MI, and patterns for patients

experiencing significant chest pain six months before the MI. For cost bucket 3, we will discuss patterns for patients with anemia and patterns for patients with cerebrovascular disease. We will also discuss patterns with diagnoses that are more commonly associated with MI; specifically diabetes, hypertension, and hyperlipidemia diagnoses. These patterns appear consistently throughout all of the cost buckets. Finally, we will discuss the clusters that have strong prediction accuracy but consist of a larger number of members with no MI.

### **5.1 Pattern of gradually increasing occurrence of COPD**

Cluster 7 from the spectral model in cost bucket 2, in Figure 5.1 below, displays a pattern where patients regularly visit the doctor for COPD. Since the disparity between the number of members who experience MI in the target period and the members who do not is moderate for Cluster 7, (see Table 6 from Section 4) we look at the “yes MI” and “no MI” members separately. However, we find that the diagnosis patterns of “yes MI” and “no MI” patients in this cluster differ significantly. Therefore, to find the unique diagnostic characteristics that “yes MI” members have in this cluster we only analyze at the “yes MI” pattern. In Figure 5.1 below, the maroon line depicts the average number of doctor visits for each diagnosis for members with MI in Cluster 7 while the blue columns are the average number of doctor visits for each diagnosis for the entire population in cost bucket 2.

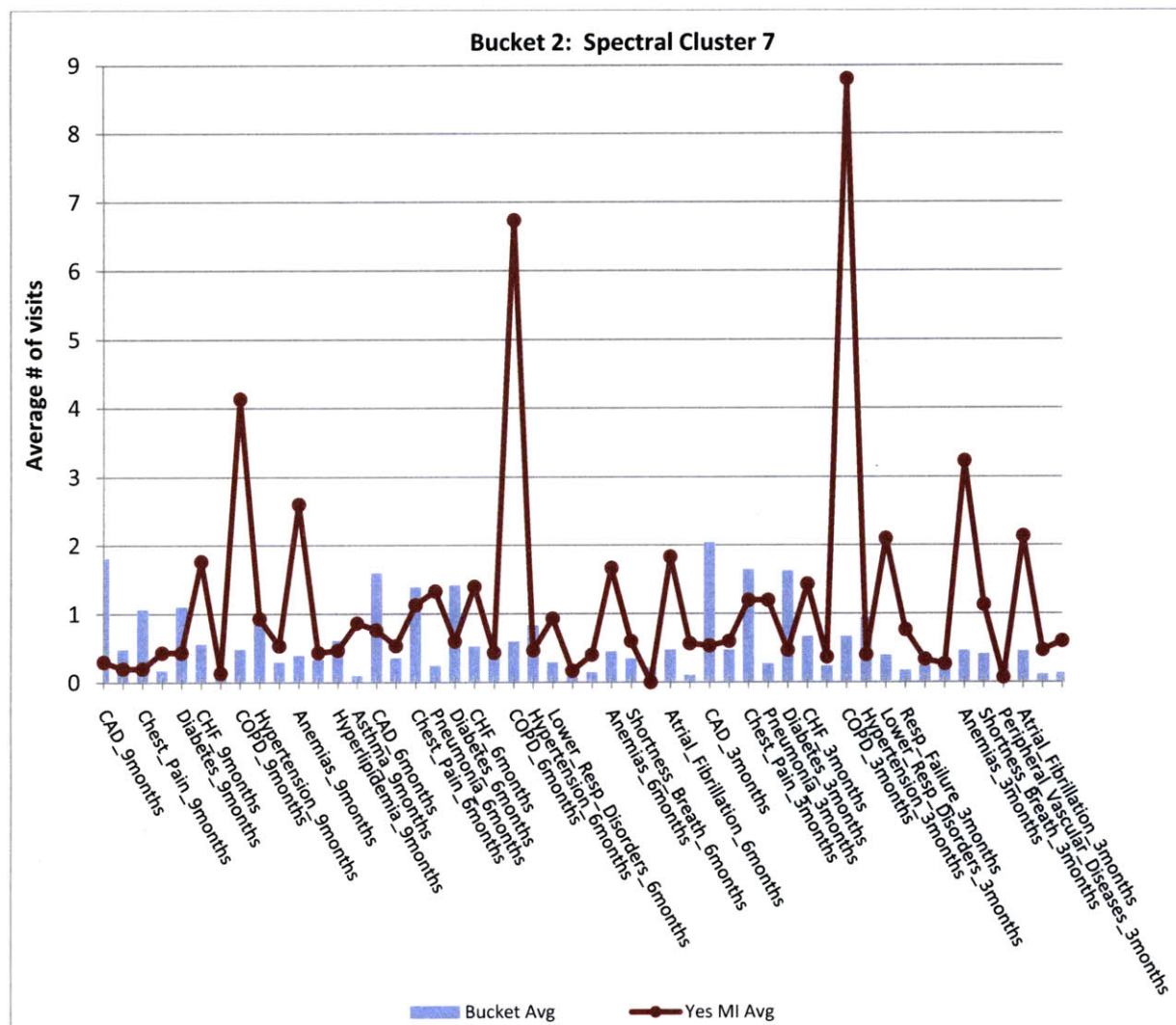


Figure 5.1: Cluster 7 in the spectral clustering model from bucket 2.

The maroon colored line in Figure 5.1 above shows that members who experience MI in the target period have an average of 4 visits to the doctor for COPD nine months before the MI, an average of 7 visits for COPD six months before the MI, and an average of 9 visits three months before. The average numbers of doctor visits for COPD are significantly larger than the average number of visits for the overall bucket 2 population. This pattern also shows that members experience Congestive Heart Failure (CHF) and anemia nine months before the MI and pneumonia, CHF, and lower respiratory disorders both three and six months before the MI. Table 5.1 below shows that compared to the cost bucket 2 population, the percentage of members with COPD three, six, and nine months before the MI is much higher in Cluster 7. The percentage with pneumonia and lower respiratory disorders is higher as well. Although pneumonia is not usually associated with myocardial infarction, information from Massachusetts

General Hospital asserts that warning signs of a heart attack can often be confused with indigestion, pneumonia, pleurisy, or other disorders [24].

COST BUCKET 2: Spectral Clustering			
FREQUENCY		Diagnosis	
Cost Bucket 2	Cluster 7		
23%	20%	9 months	Coronary Artery Disease
14%	11%		Chest Pain
6%	9%		Congestive Heart Failure
8%	56%		Chronic Obstructive Pulmonary Disease
26%	16%		Hypertension
5%	18%		Anemias
23%	27%	6 months	Coronary Artery Disease
15%	15%		Chest Pain
4%	18%		Pneumonia
24%	22%		Diabetes
7%	15%		Congestive Heart Failure
9%	64%		Chronic Obstructive Pulmonary Disease
27%	29%		Hypertension
8%	24%		Lower Respiratory Disorders
6%	25%		Anemias
10%	24%		Shortness of Breath
5%	7%		Atrial Fibrillation
27%	25%	3 months	Coronary Artery Disease
18%	13%		Chest Pain
5%	15%		Pneumonia
26%	22%		Diabetes
8%	20%		Congestive Heart Failure
10%	67%		Chronic Obstructive Pulmonary Disease
26%	31%		Hypertension
9%	25%		Lower Respiratory Disorders
6%	24%		Anemias
11%	27%		Shortness of Breath
5%	7%		Atrial Fibrillation

Table 5.1: Diagnoses that distinguish Cluster 7 from the cost bucket 2 population average. The first two columns show the percentage of members who have the listed diagnosis.

Cluster 8 from the k-means model also has a similar temporal pattern of COPD diagnoses as in the spectral Cluster 7 above. In Figure 5.2 below we find that the most significant diagnosis members in Cluster 8 experience nine, six, and three months before the MI is COPD. Table 5.2 shows that a very large percentage of members in this cluster have COPD nine, six, and three months before the MI. In order to better show that this subgroup of members with COPD have an increased risk of MI, we graph the Cluster 8 pattern against members from the bucket 2 population having a COPD diagnosis. Therefore, the blue columns in Figure 5.2 are the average number of doctor visits for each diagnosis only for members in the cost bucket 2 population that have at least one visit to the doctor for COPD. The significant difference in the average number of doctor visits for COPD between Cluster 8 members and the conditional bucket 2 population shows that members with this particular pattern and magnitude of COPD have an increased risk of MI.

Figure 5.2 also shows that members in Cluster 8 visit the doctor an average of 10 times for CAD nine months before the MI. This is different from what we found in Cluster 7 above, but Table 5.2 below shows that only 33% of members in this cluster actually visit the doctor for CAD. Therefore, this small percentage of members is unrepresentative of the Cluster 8 population and the result for CAD nine months before the MI is skewed. In addition to the COPD diagnosis, members in this cluster develop CHF, respiratory failure, and lower respiratory disorders six and three months before the MI and then peripheral vascular diseases and diseases of pulmonary circulation three months before the MI. The patterns of COPD found in these clusters are supported by an independent study done on the management of COPD patients. Researchers in this study found that in patients with acute myocardial infarction, the incidence of COPD was approximately 50% higher than in the general population [26].



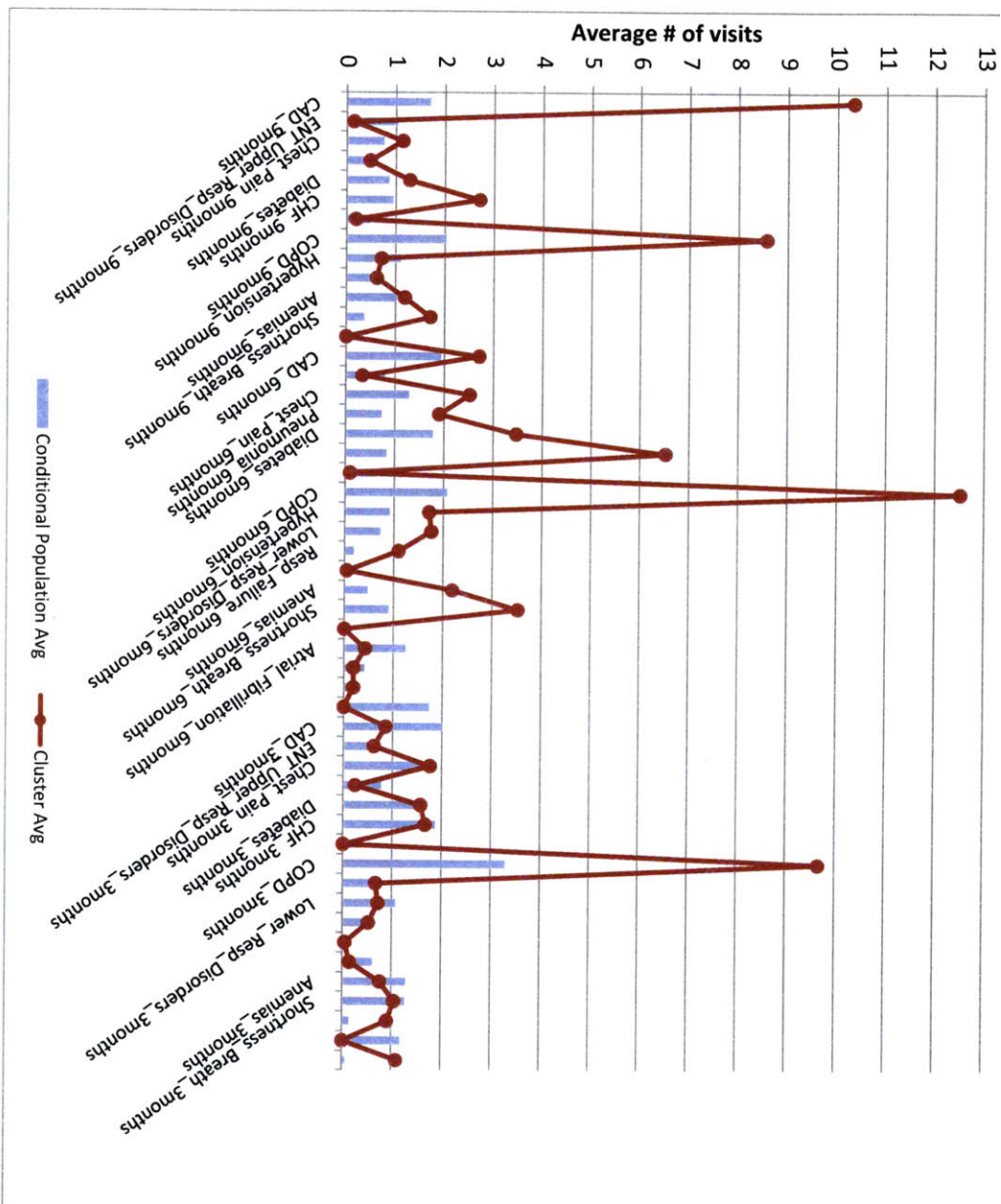


Figure 5.2: Cluster 8 in the k-means clustering model from bucket 2.

COST BUCKET 2: K-means Clustering			
FREQUENCY		Diagnosis	
Cost Bucket 2	Cluster 8		
23%	33%	9 months	Coronary Artery Disease
24%	33%		Diabetes
6%	14%		Congestive Heart Failure
8%	76%		Chronic Obstructive Pulmonary Disease
5%	14%		Anemias
8%	52%		Shortness of breath
16%	19%		Hyperlipidemia
2%	14%		Asthma
23%	52%	6 months	Coronary Artery Disease
15%	33%		Chest Pain
4%	33%		Pneumonia
24%	38%		Diabetes
7%	29%		Congestive Heart Failure
9%	90%		Chronic Obstructive Pulmonary Disease
27%	43%		Hypertension
8%	33%		Lower Respiratory Disorders
2%	24%		Respiratory Failure
6%	24%		Anemias
10%	52%		Shortness of breath
18%	24%	3 months	Chest Pain
26%	24%		Diabetes
8%	19%		Congestive Heart Failure
10%	81%		Chronic Obstructive Pulmonary Disease
26%	24%		Hypertension
9%	24%		Lower Respiratory Disorders
5%	19%		Respiratory Failure
11%	33%		Shortness of breath
5%	19%		Peripheral Vascular Diseases
1%	19%		Diseases Pulmonary Circulation

Table 5.2: Diagnoses that distinguish Cluster 8 from the cost bucket 2 population average. The first two columns show the percentage of members who have the listed diagnosis.

## 5.2 Pattern of Deteriorating Health

Cluster 1 from the k-means model in cost bucket 2 not only has a high prediction rate, but also possesses a temporal pattern of an interesting combination of diagnoses that is much different

than the general population of members in cost bucket 2. Figure 5.3 provides a graphical representation of this pattern. The red line depicts the average number of doctor visits for each diagnosis for members in Cluster 1. The blue columns are the average number of doctor visits for each diagnosis for the entire population in cost bucket 2.

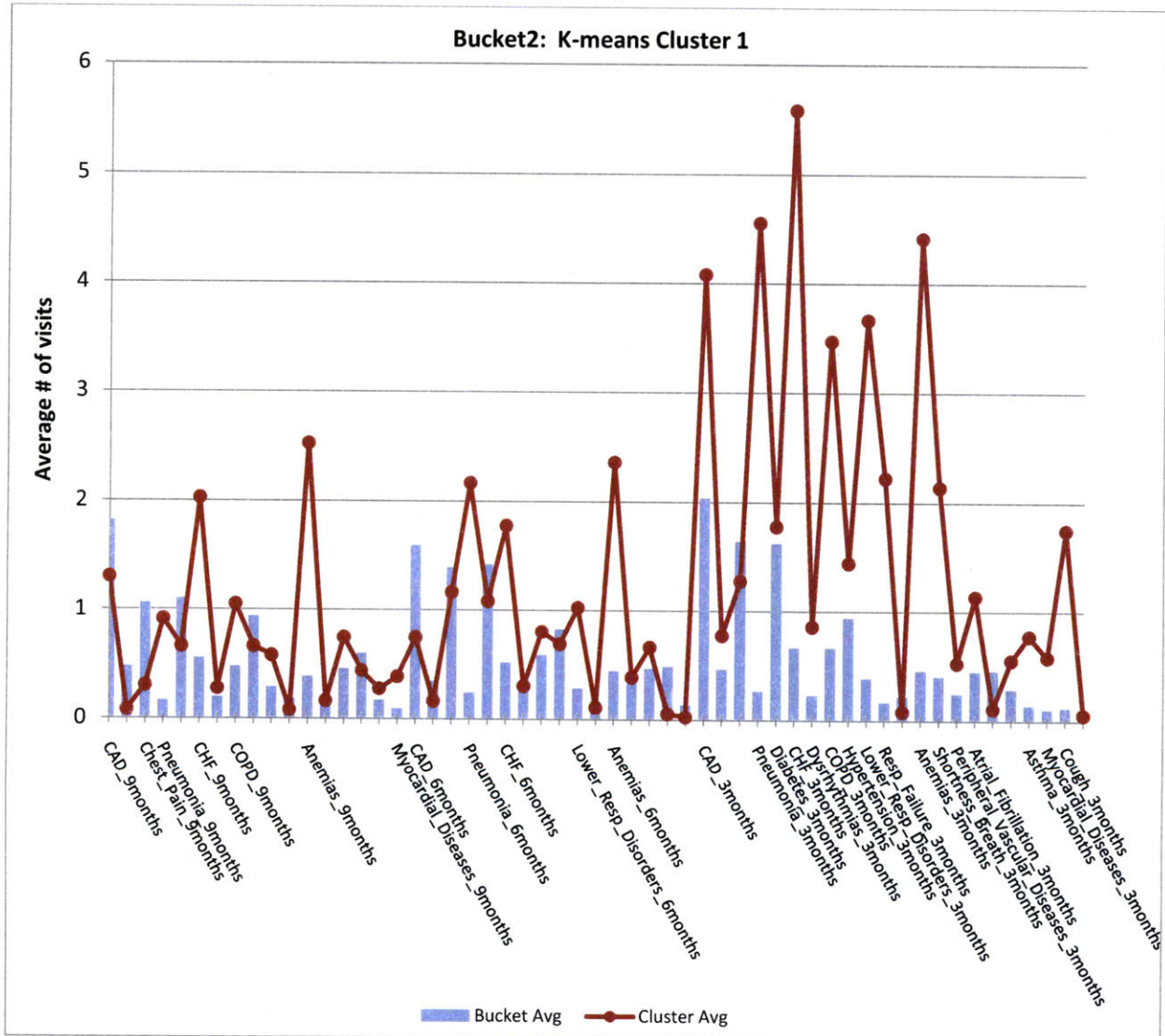


Figure 5.3: Cluster 1 in the k-means clustering model from bucket 2.

The red line in Figure 5.3 shows the progressive deterioration of health for the members in Cluster 1. Leading up to the heart attack in the target period, members in this cluster go to the doctor an average of 2 times with CHF and anemia nine months before the MI. Six months out from the MI, members visit the doctor an average of 2 times with pneumonia, CHF, and anemia.



Three months before the MI, the health condition of members becomes worse and it is the combination of diagnoses that makes these members unique. Members visit the doctor an average of 4 times with CAD, almost 5 times with pneumonia, anemia, and shortness of breath, almost 6 times with CHF, almost 4 times with COPD, lower respiratory disorders, and respiratory failure, and 2 times with a cough. Many of these diagnoses that the members experience three months before the MI can be associated with a bad episode of the flu and because of these misdiagnoses, some cardiologists believe that the signs and symptoms of a heart attack are often missed [25]. They say that many times a patient's routine Electrocardiogram (EKG) will show signs of a recent heart attack and the patient has no idea a heart attack occurred. Instead, the patient thinks they have a really bad flu [25]. Table 5.3 below shows that in particular, the percentages of members in Cluster 1 with pneumonia, lower respiratory disorders, shortness of breath, and a cough three months before the MI are much greater than the overall cost bucket 2 population percentages.

COST BUCKET 2: K-means Clustering			
FREQUENCY		Diagnosis	
Cost Bucket 2	Cluster 1		
23%	19%	9 months	Coronary Artery Disease
3%	11%		Pneumonia
6%	22%		Congestive Heart Failure
8%	22%		Chronic Obstructive Pulmonary Disease
5%	6%		Anemias
3%	17%		Myocardial Diseases
23%	28%	6 months	Coronary Artery Disease
4%	19%		Pneumonia
7%	25%		Congestive Heart Failure
8%	31%		Lower Respiratory Disorders
6%	11%		Anemias
27%	39%	3 months	Coronary Artery Disease
5%	56%		Pneumonia
26%	33%		Diabetes
8%	53%		Congestive Heart Failure
9%	36%		Dysrhythmias
10%	44%		Chronic Obstructive Pulmonary Disease
26%	33%		Hypertension
9%	53%		Lower Respiratory Disorders
5%	36%		Respiratory Failure
6%	22%		Anemias
11%	39%		Shortness of breath
5%	14%		Peripheral Vascular Diseases
5%	19%		Atrial Fibrillation
2%	8%		Asthma
5%	22%		Myocardial Diseases
3%	33%		Cough

Table 5.3: Diagnoses that distinguish Cluster 1 from the cost bucket 2 population average. The first two columns show the percentage of members who have the listed diagnosis.

### 5.3 Pattern of significant occurrence of Chest Pain 3 Months before MI

Cluster 1 from the spectral model in cost bucket 2, shown in Figure 5.4, depicts a pattern of numerous doctors visits for chest pain three months before the MI. The red line depicts the average number of doctor visits for each diagnosis for members in Cluster 1. The blue columns

are the average number of doctor visits for each diagnosis for the entire population in cost bucket 2.

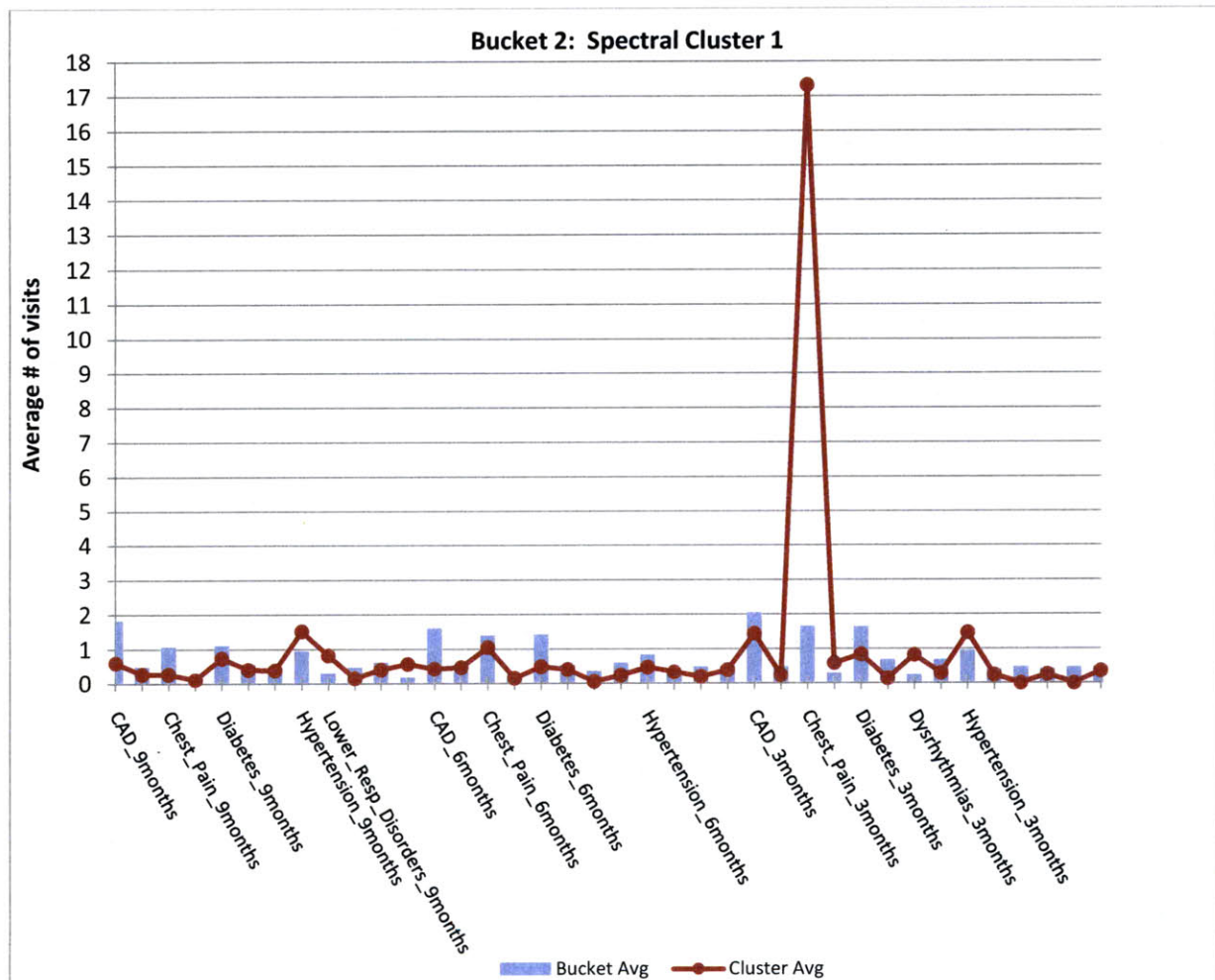


Figure 5.4: Cluster 1 in the spectral clustering model from bucket 2.

Members in Cluster 1 above have relatively low number of doctor visits early on in their medical history. However, three months before the MI, the average number of visits to the doctor with a chest pain diagnosis is almost 18. Table 5.4 below shows that 100% of the members in this cluster had a chest pain diagnosis three months before the MI. We also find that, compared to the bucket 2 population average, a large percentage of members had visits to the doctor for CAD and dysrhythmias three months before the MI. Although this pattern of diagnoses is something

we would expect for a member that is about to experience MI, it is interesting and useful to find this pattern in a systematic way.

COST BUCKET 2: Spectral Clustering			
FREQUENCY		Diagnosis	
Cost Bucket 2	Cluster 1		
23%	14%	9 months	Coronary Artery Disease
26%	24%		Hypertension
8%	14%		Lower Respiratory Disorders
23%	18%	6 months	Coronary Artery Disease
15%	18%		Chest Pain
27%	18%		Hypertension
27%	43%	3 months	Coronary Artery Disease
18%	100%		Chest Pain
9%	22%		Dysrhythmias
26%	27%		Hypertension

Table 5.4: Diagnoses that distinguish Cluster 1 from the cost bucket 2 population average. The first two columns show the percentage of members who have the listed diagnosis.

Cluster 7 in the k-means model, shown in Figure 5.5 below, also has a pattern of chest pain 3 months before the MI similar to spectral cluster 1 above. In order to better show that this subgroup of members with chest pain have an increased risk of MI, we graph the Cluster 7 pattern against members from the bucket 2 population having a chest pain diagnosis three months before the target MI period. Therefore, the blue columns in Figure 5.5 are the average number of doctor visits for each diagnosis only for members in the cost bucket 2 population that have at least one visit to the doctor for chest pain three months before the target period.



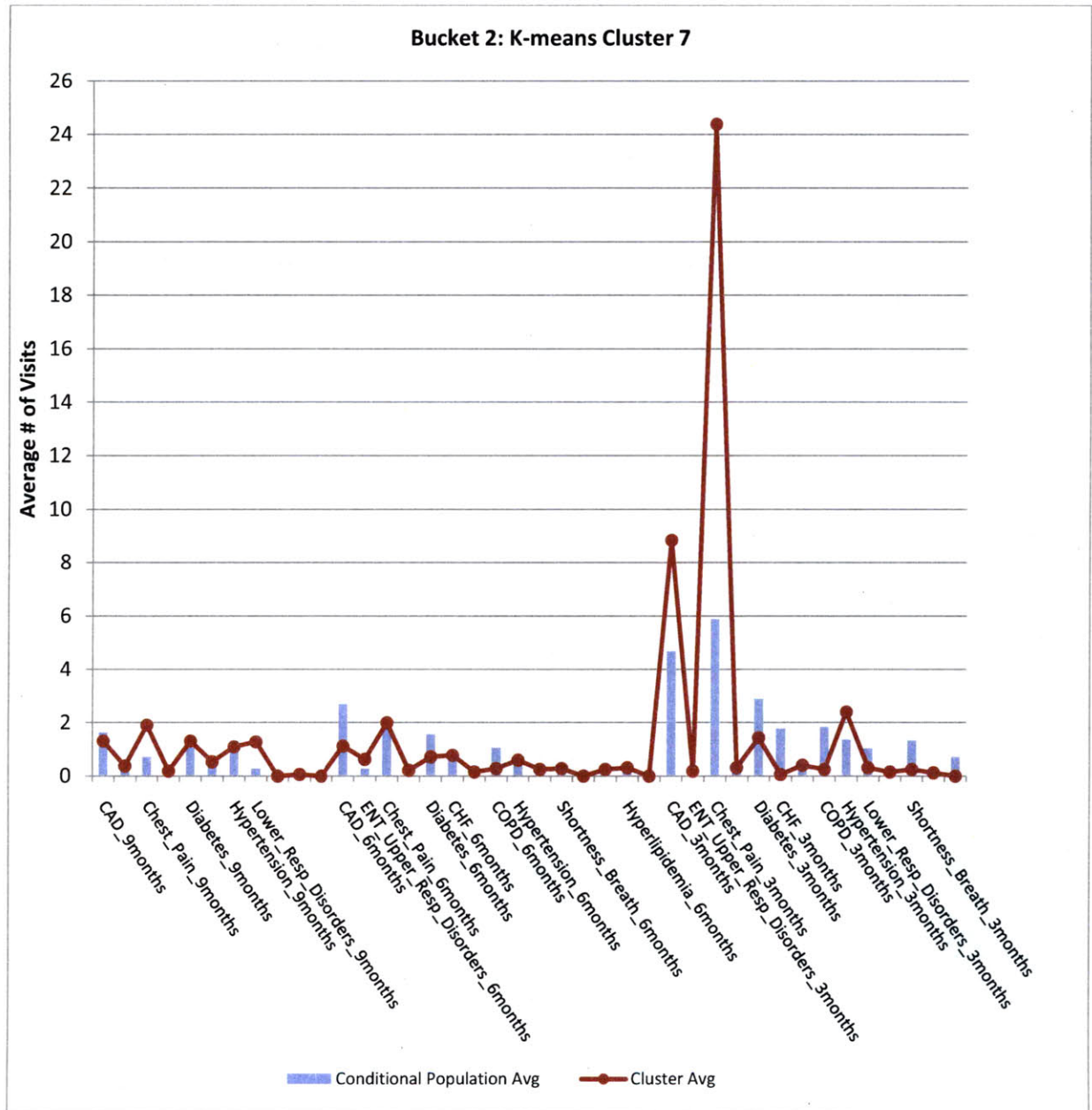


Figure 5.5: Cluster 7 in the k-means clustering model from bucket 2.

The red line in Figure 5.5 above shows that members in Cluster 7 have a moderate history of visits to the doctor six and nine months before the MI. The temporal pattern of this cluster only differs from the cost bucket 2 averages in a few places: nine months before the MI, members visit the doctor an average of 2 times for chest pain and about once for lower respiratory disorders and six months before the MI, members visit the doctor an average of 2 times for chest pain. Then, three months before the MI, members visit the doctor an average of 9 times for

CAD, almost 25 times for chest pain, and 2 times for hypertension. The significant difference in the average number of doctor visits for chest pain three months before the MI between Cluster 7 members and the conditional bucket 2 population shows that members with this particular pattern and magnitude of chest pain have an increased risk of MI.

Table 5.5 below shows that while a higher percentage of members in Cluster 7 have chest pain six and nine months before the MI (when compared to cost bucket 2 percentages), three months before the MI 100% of the members have chest pain. Additionally, the percentage of members in Cluster 7 that have lower respiratory disorders nine and three months before the MI are higher than the overall cost bucket 2 percentages.

COST BUCKET 2: K-means Clustering			
FREQUENCY		Diagnosis	
Cost Bucket 2	Cluster 7		
14%	25%	9 months	Chest Pain
8%	13%		Chronic Obstructive Pulmonary Disease
26%	19%		Hypertension
8%	19%		Lower Respiratory Disorders
23%	22%	6 months	Coronary Artery Disease
11%	16%		Ear, Nose, Throat & Upper Respiratory Disorders
15%	25%		Chest Pain
7%	6%		Congestive Heart Failure
27%	22%		Hypertension
10%	19%		Shortness of breath
27%	56%	3 months	Coronary Artery Disease
18%	100%		Chest Pain
26%	25%		Diabetes
8%	6%		Congestive Heart Failure
9%	22%		Dysrhythmias
26%	31%		Hypertension
9%	16%		Lower Respiratory Disorders
11%	19%		Shortness of breath

Table 5.5: Diagnoses that distinguish Cluster 7 from the cost bucket 2 population average. The first two columns show the percentage of members who have the listed diagnosis.

## **5.4 Pattern of significant occurrence of CAD 3 Months before MI**

Cluster 6 from the spectral model in cost bucket 2, shown below in Figure 5.6, depicts a pattern of numerous doctors visits for CAD three months before the MI. Since the disparity between the number of members who experience MI in the target period and the members who do not is moderate for Cluster 6, we look at the “yes MI” and “no MI” members separately. This helps to determine what unique diagnostic characteristics the “yes MI” members have and why we predict MI so well for this cluster. In Figure 5.6, the maroon line depicts the average number of doctor visits for each diagnosis for members with MI in Cluster 6 and the green line depicts average number of doctor visits for each diagnosis for members with no MI in Cluster 6. The blue columns are the average number of doctor visits for each diagnosis for the entire population in cost bucket 2.

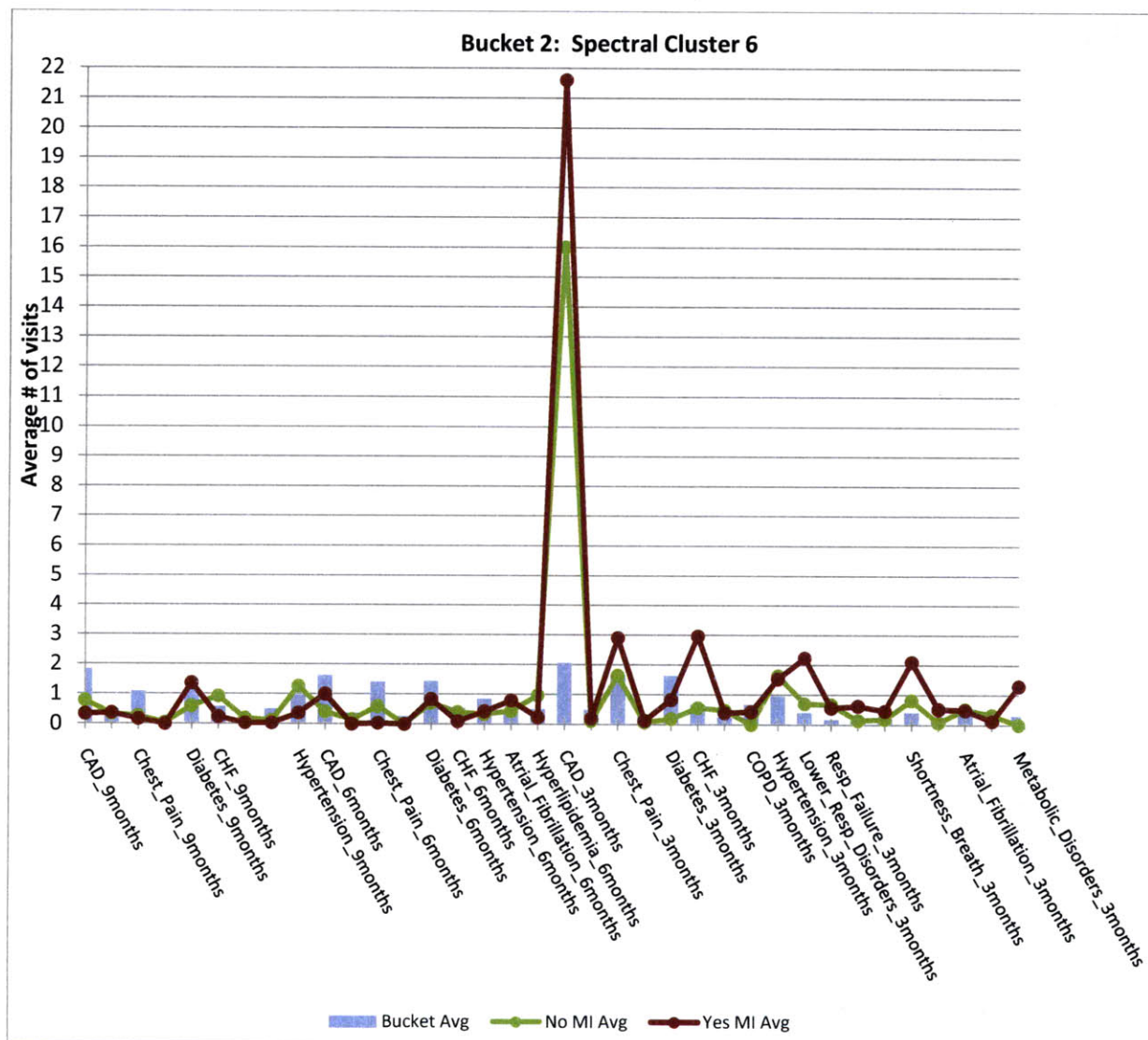


Figure 5.6: Cluster 6 in the spectral clustering model from bucket 2.

Members in Cluster 6 have a relatively low number of doctor visits until 3 months before the MI target period. The maroon colored line in Figure 5.6 above shows that members who have MI in the target period visit the doctor an average of 22 times with the CAD diagnosis. The line also shows that members with MI in the target period have more visits to the doctor for chest pain, Congestive Heart Failure (CHF), lower respiratory disorders, and shortness of breath. Table 5.6 below shows that compared to the entire cost bucket 2 population, the percentage of members in Cluster 6 with these diagnoses three months before the MI is much higher.



COST BUCKET 2: Spectral Clustering			
FREQUENCY		Diagnosis	
Cost Bucket 2	Cluster 6		
23%	28%	9 months	Coronary Artery Disease
14%	7%		Chest Pain
24%	14%		Diabetes
26%	35%		Hypertension
23%	25%	6 months	Coronary Artery Disease
15%	7%		Chest Pain
24%	19%		Diabetes
27%	19%		Hypertension
27%	100%	3 months	Coronary Artery Disease
18%	39%		Chest Pain
26%	21%		Diabetes
8%	16%		Congestive Heart Failure
26%	37%		Hypertension
9%	23%		Lower Respiratory Disorders
11%	21%		Shortness of Breath

Table 5.6: Diagnoses that distinguish Cluster 6 from the cost bucket 2 population average. The first two columns show the percentage of members who have the listed diagnosis.

## 5.5 Pattern of significant occurrence of Chest Pain 6 Months before MI

Cluster 8 from the spectral model in cost bucket 2, shown in Figure 5.7 below, consists of members who visit the doctor numerous times for chest pain six months before the MI. The red line depicts the average number of doctor visits for each diagnosis for members in Cluster 8. The blue columns are the average number of doctor visits for each diagnosis for the entire population in cost bucket 2.

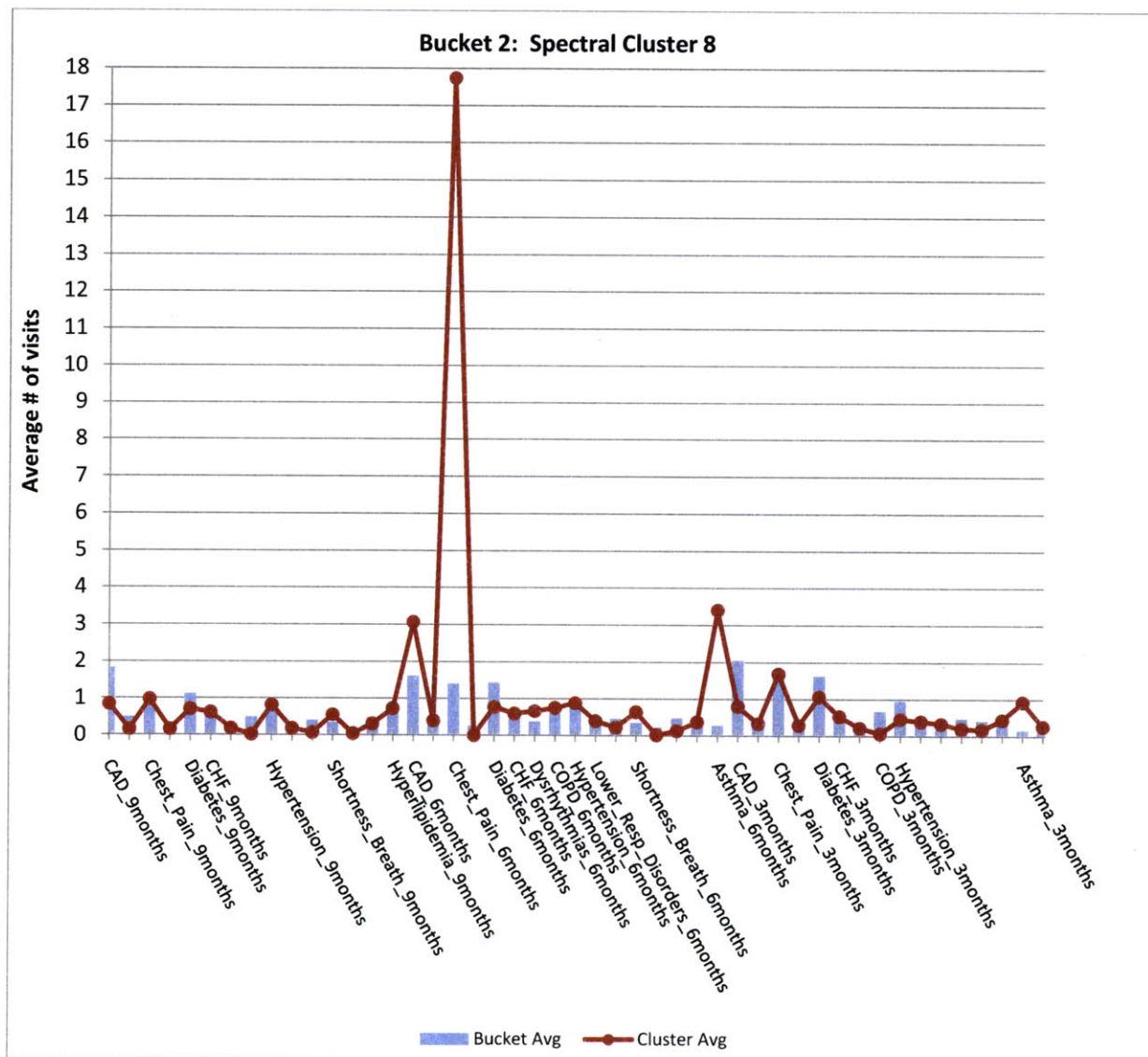


Figure 5.7: Cluster 8 in the spectral clustering model from bucket 2.

The red line in Figure 5.7 above shows that members from Cluster 8 visited the doctor an average of 18 times for chest pain six months before the MI. Although chest pain is the key warning sign of an immediate heart attack, the timing of this diagnosis is surprising. We also find that members in Cluster 8 have a greater number of visits to the doctor for asthma than normal. Table 5.7 below shows that 100% of the members in Cluster 8 visit the doctor for chest pain 6 months before the MI and a high percentage of members have CAD, lower respiratory disorders, and shortness of breath 6 months before the MI.

COST BUCKET 2: Spectral Clustering			
FREQUENCY		Diagnosis	
Cost Bucket 2	Cluster 8		
23%	16%	9 months	Coronary Artery Disease
14%	13%		Chest Pain
26%	27%		Hypertension
16%	18%		Hyperlipidemia
23%	42%	6 months	Coronary Artery Disease
15%	100%		Chest Pain
27%	33%		Hypertension
8%	18%		Lower Respiratory Disorders
10%	22%		Shortness of Breath
2%	4%		Asthma
27%	36%	3 months	Coronary Artery Disease
18%	27%		Chest Pain
26%	24%		Diabetes
8%	9%		Congestive Heart Failure
26%	24%		Hypertension
2%	7%		Asthma

Table 5.7: Diagnoses that distinguish cluster 8 from the cost bucket 2 population average. The first two columns show the percentage of members who have the listed diagnosis.

## 5.6 Pattern of gradually increasing occurrence of Anemia

Figure 5.8 below is a graphical representation of the temporal diagnosis pattern of anemia in Cluster 4 from cost bucket 3. In order to better show that this subgroup of members with anemia have an increased risk of MI, we graph the Cluster 4 pattern against members from the bucket 2 population having an anemia diagnosis. Therefore, the blue columns in Figure 5.8 are the average number of doctor visits for each diagnosis only for members in the cost bucket 2 population that have at least one visit to the doctor for anemia.



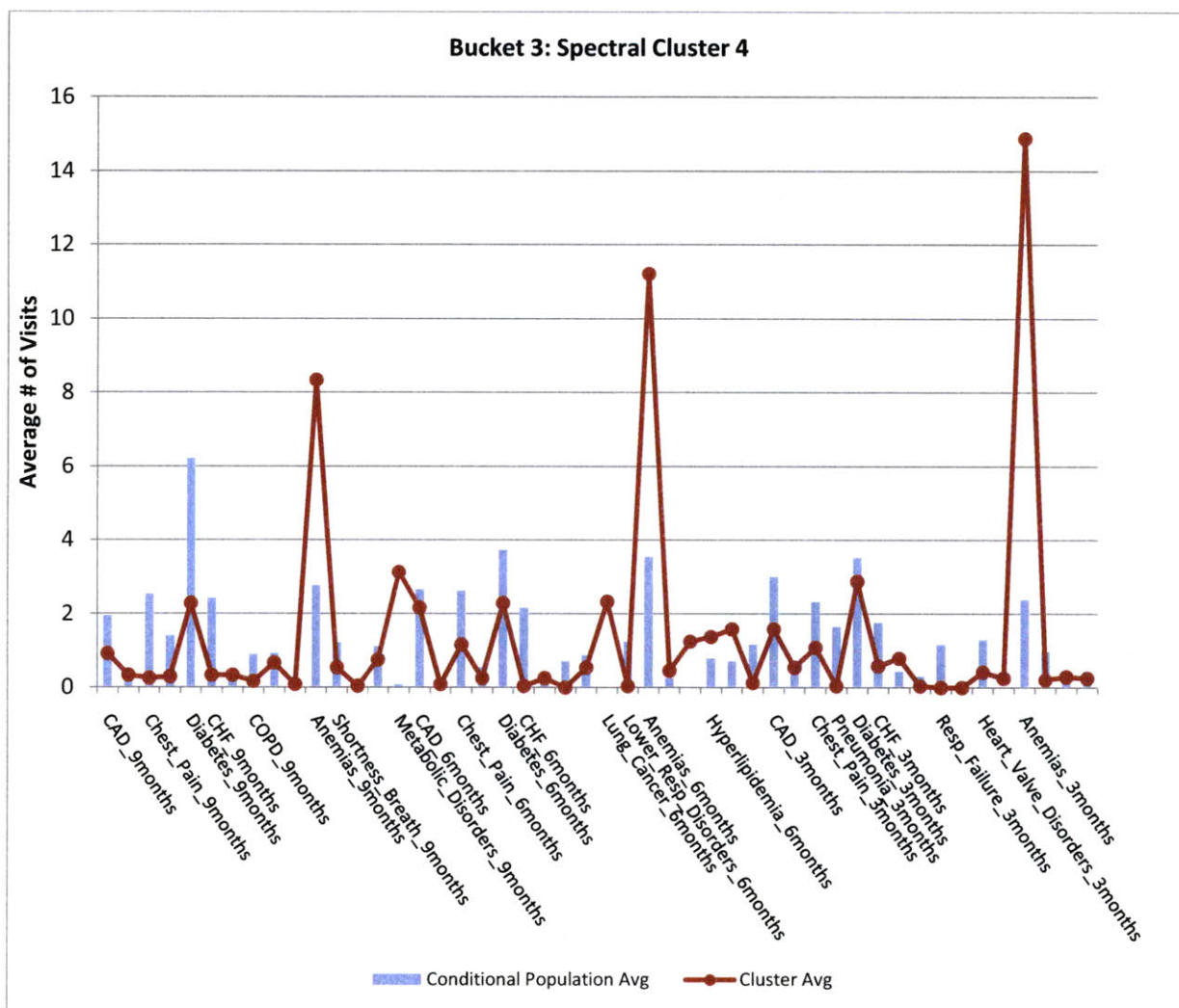


Figure 5.8: Cluster 4 in the spectral clustering model from bucket 3.

The red line shows that members in Cluster 4 increasingly occurrence visit the doctor for anemia from nine to three months before the MI. Nine months before the MI, members have on average 9 visits to the doctor for anemia. This increases to an average of 11 visits six months before the MI and then an average of 15 visits three months before the MI. Table 5.8 shows that a very large percentage of members in this cluster have anemia nine, six, and three months before the MI. In addition, the significant difference in the average number of doctor visits for anemia between Cluster 4 members and the conditional bucket 2 population shows that members with this particular pattern and magnitude of anemia have an increased risk of MI. According to the International Academy of Cardiology, a study done at Brigham and Women's Hospital found that "anemia has been shown to significantly decrease oxygen delivery to the myocardium and

increase the myocardial oxygen demand”. The increase in blood viscosity can lead to a decrease in coronary blood flow which can ultimately lead to thrombosis and MI. [27].

COST BUCKET 3: Spectral Clustering			
FREQUENCY		Diagnosis	
Cost Bucket 3	Cluster 4		
26%	25%	9 months	Coronary Artery Disease
14%	8%		Chest Pain
31%	46%		Diabetes
4%	17%		Respiratory Failure
8%	46%		Anemias
4%	17%		Peripheral Vascular Diseases
3%	17%		Metabolic Disorders
28%	29%	6 months	Coronary Artery Disease
21%	21%		Chest Pain
29%	46%		Diabetes
1%	4%		Lung Cancer
10%	79%		Anemias
3%	8%		Atrial Fibrillation
16%	13%		Hyperlipidemia
6%	17%		Metabolic Disorders
32%	46%	3 months	Coronary Artery Disease
19%	21%		Chest Pain
28%	42%		Diabetes
11%	79%		Anemias
6%	13%		Metabolic Disorders

Table 5.8: Diagnoses that distinguish Cluster 4 from the cost bucket 3 population average. The first two columns show the percentage of members who have the listed diagnosis.

## 5.7 Pattern of gradually increasing occurrence of Cerebrovascular Disease

Further analysis of Cluster 5 from cost bucket 3 shows that relative to the general population of cost bucket 3, it has a unique and interesting temporal pattern of cerebrovascular disease that leads to MI. Figure 5.9 provides a graphical representation of this pattern.

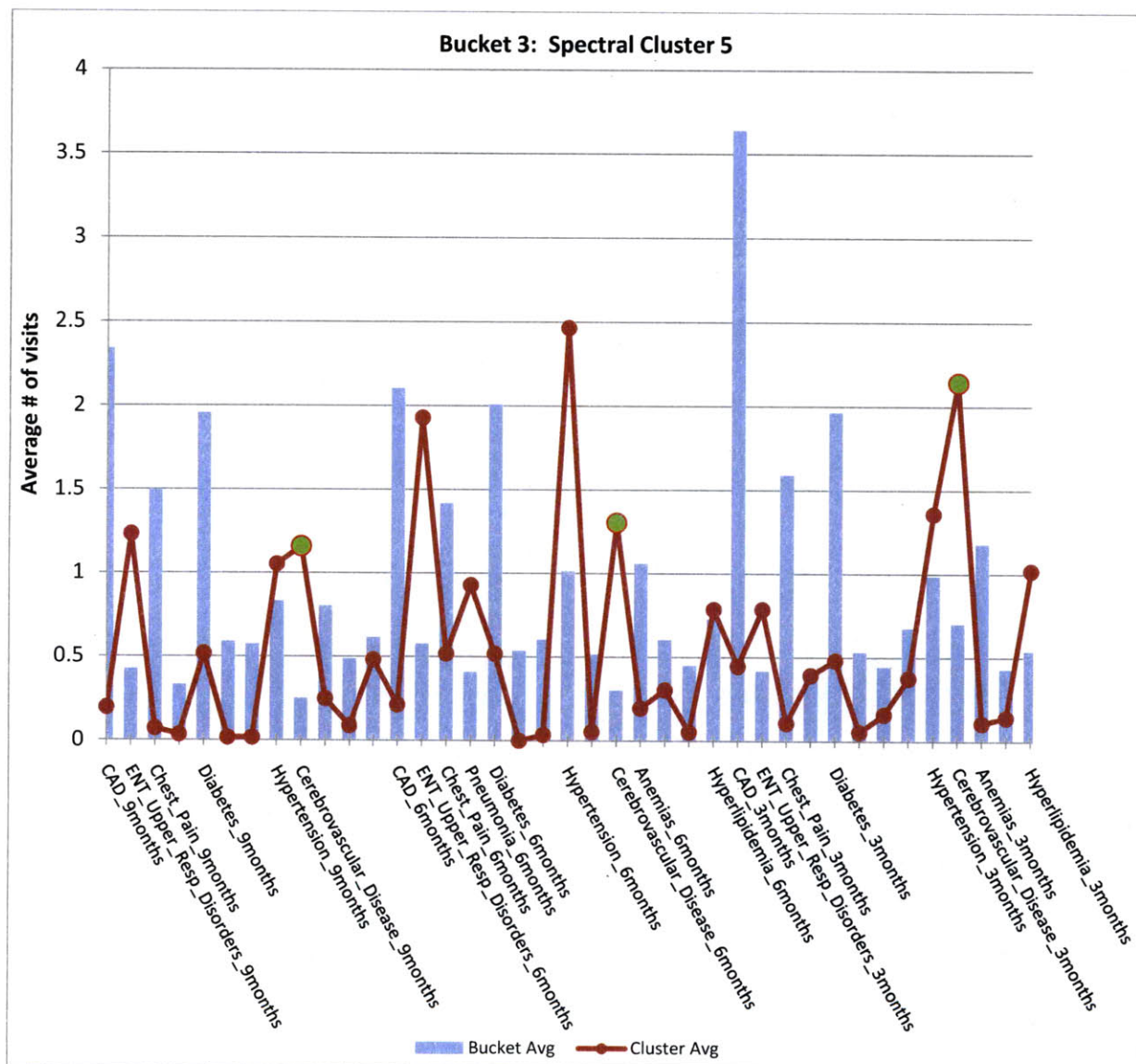


Figure 5.9: Cluster 5 in the spectral clustering model from bucket 3.

From the red line in Figure 5.9 above we observe that members from Cluster 5 have a reoccurring pattern of doctor visits that include diagnoses of hypertension and cerebrovascular disease. To highlight the significant disparity between the average number of doctor visits for cerebrovascular disease for members in Cluster 5 and the overall cost bucket 2 population, we color the cerebrovascular disease diagnosis in green on the graph above. The pattern also shows that members of this cluster have more health complications occurring six months before the MI. Nine months before the MI, members only have visits to the doctor for ENT and upper respiratory disorders, hypertension, and cerebrovascular disease. Six months from the MI, members have a more complicated combination of diagnoses to include ENT and upper



respiratory disorders, chest pain, pneumonia, hypertension, and cerebrovascular disease. Three months before the MI members have hypertension, cerebrovascular disease, and hyperlipidemia. The diagnosis combinations that occur in this temporal pattern are interesting because they do not fall under the typical signs and symptoms of MI. However, research from the European Society of Cardiology has shown that myocardial infarctions are common in patients with cerebrovascular disease and despite conventional diagnostic procedures they often pass undiscovered [28]. Table 5.9 below shows that the percentage of members with cerebrovascular disease in Cluster 5 is significantly larger than the percentage of members with cerebrovascular disease in the overall cost bucket 3 population.

COST BUCKET 3: Spectral Clustering			
FREQUENCY		Diagnosis	
Cost Bucket 3	Cluster 5		
9%	10%	9 months	Ear, Nose, Throat & Upper Respiratory Disorders
25%	21%		Hypertension
5%	11%		Cerebrovascular Disease
28%	13%	6 months	Coronary Artery Disease
13%	18%		Ear, Nose, Throat & Upper Respiratory Disorders
6%	4%		Pneumonia
29%	16%		Diabetes
27%	34%		Hypertension
6%	18%		Cerebrovascular Disease
10%	4%		Anemias
16%	14%	3 months	Hyperlipidemia
32%	20%		Coronary Artery Disease
12%	16%		Ear, Nose, Throat & Upper Respiratory Disorders
28%	11%		Diabetes
28%	30%		Hypertension
7%	9%		Cerebrovascular Disease
14%	13%		Hyperlipidemia

Table 5.9: Diagnoses that distinguish Cluster 5 from the cost bucket 3 population average. The first two columns show the percentage of members who have the listed diagnosis.

## 5.8 Patterns associated with Diabetes, Hypertension, and Hyperlipidemia

In addition to the clusters described above, the algorithms found other unique clusters with interesting diagnostic temporal patterns that lead to MI. Moreover, the diagnostic characteristics of the members within these clusters are well-recognized in the medical community and society.

For example, Cluster 10 from the spectral model in cost bucket 3 exhibits a temporal diagnostic pattern of diabetes (Figure 5.10). It is commonly known that both types 1 and 2 diabetes are associated with accelerated atherosclerosis, one of the main causes of myocardial infarction [3].

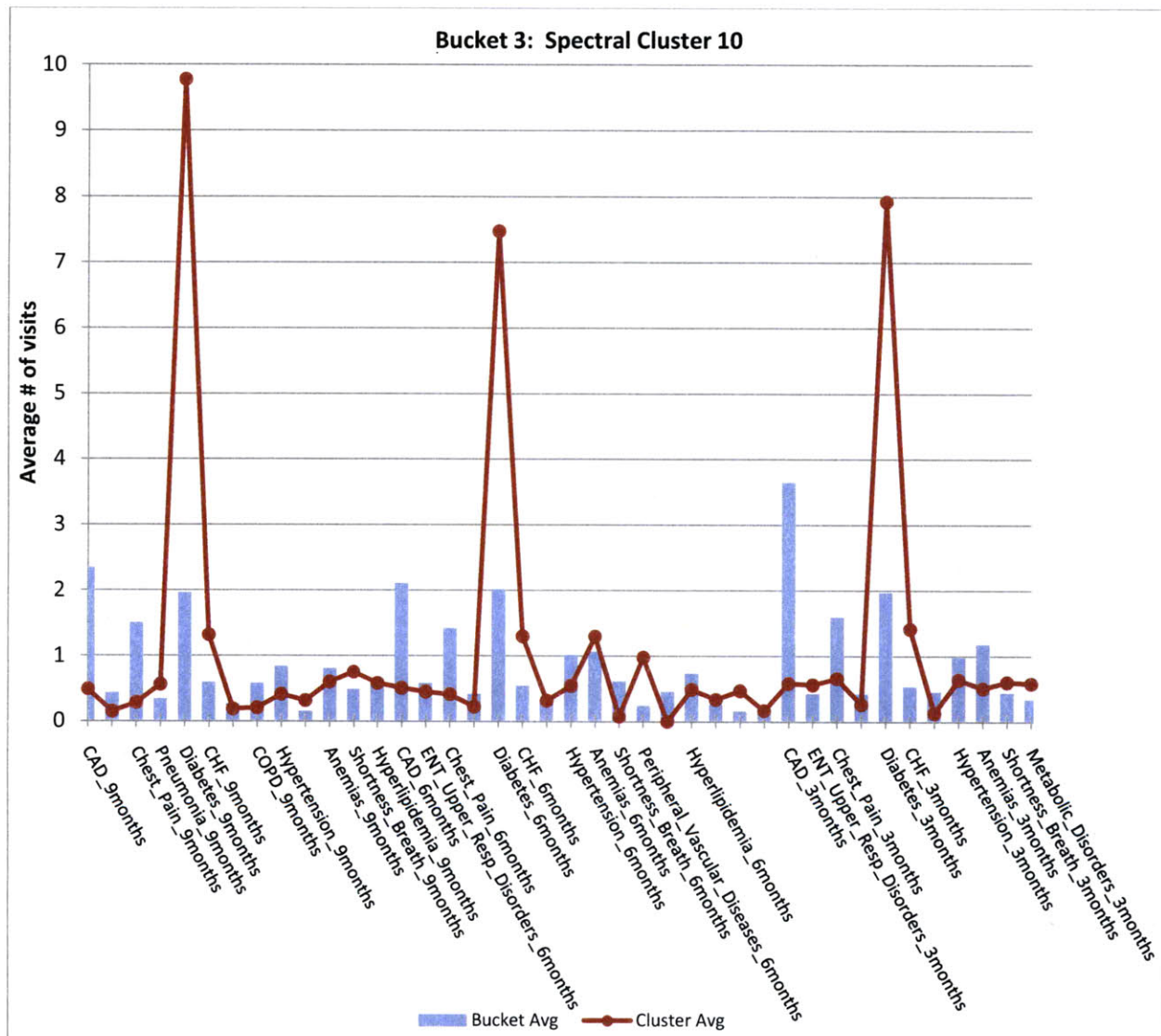


Figure 5.10: Cluster 10 in the spectral clustering model from bucket 3. The red line depicts the average number of doctor visits for each diagnosis for members in Cluster 10. The blue columns are the average number of doctor visits for each diagnosis for the entire population in bucket 3.

In Section 1, we reviewed the diagnoses that are well-known to lead to MI by the medical community and society. These diagnoses - diabetes, hypertension and hyperlipidemia -



characterize many of the patterns that consistently occur throughout all of the cost buckets and clustering models. In the spectral clustering model from cost bucket 1, Cluster 1 consists of members with hypertension. Members from Cluster 3 experience hypertension only six months before the MI and members from Cluster 4 experience hypertension only three months before the MI. Cluster 8 contains members with hyperlipidemia and in Cluster 9 members have diabetes. While using the k-means clustering algorithm in cost bucket 1, we also encountered patterns of well-known diagnoses. For example, one cluster consisted of members with hyperlipidemia, another consisted of members who have diabetes, and another contained members with hypertension. In cost bucket 2 we find the same kinds of patterns. Cluster 5 from the spectral model and Cluster 2 from the k-means model both contain members with diabetes. Cluster 10 from the spectral model and Cluster 10 from the k-means model both consist of members with hypertension. In Cluster 9 from the k-means model, members have complicated hypertension. In cost bucket 3, the diagnostic patterns are slightly more complicated because patients have worse health conditions, but we still find patterns that consist of well-known diagnoses leading to MI: the diabetes members in Cluster 10 (Figure 5.10 above), members in Cluster 7 who have hyperlipidemia and chest pain, and members in Cluster 9 who have only chest pain.

In addition to the validity of our clusters, we find that our results are also stable. The stability of the temporal patterns we found is evident because many of the diagnostic patterns occur in clusters from both algorithms. This helps to show that our results are not a coincidental product of a single clustering algorithm.

## **5.9 Patterns associated with No MI**

To enhance the analysis from above we present a few insights into the patterns associated with clusters consisting of a large number of “no MI” members. These findings can be helpful in not only showing what patterns of diagnoses are not associated with heart attacks, but also how the timing of certain diagnostic events is critical. In Figure 5.11 below, we see that members in spectral Cluster 3 from Bucket 2 visit the doctor an average of 15 times for chest pain nine months before the MI target period.

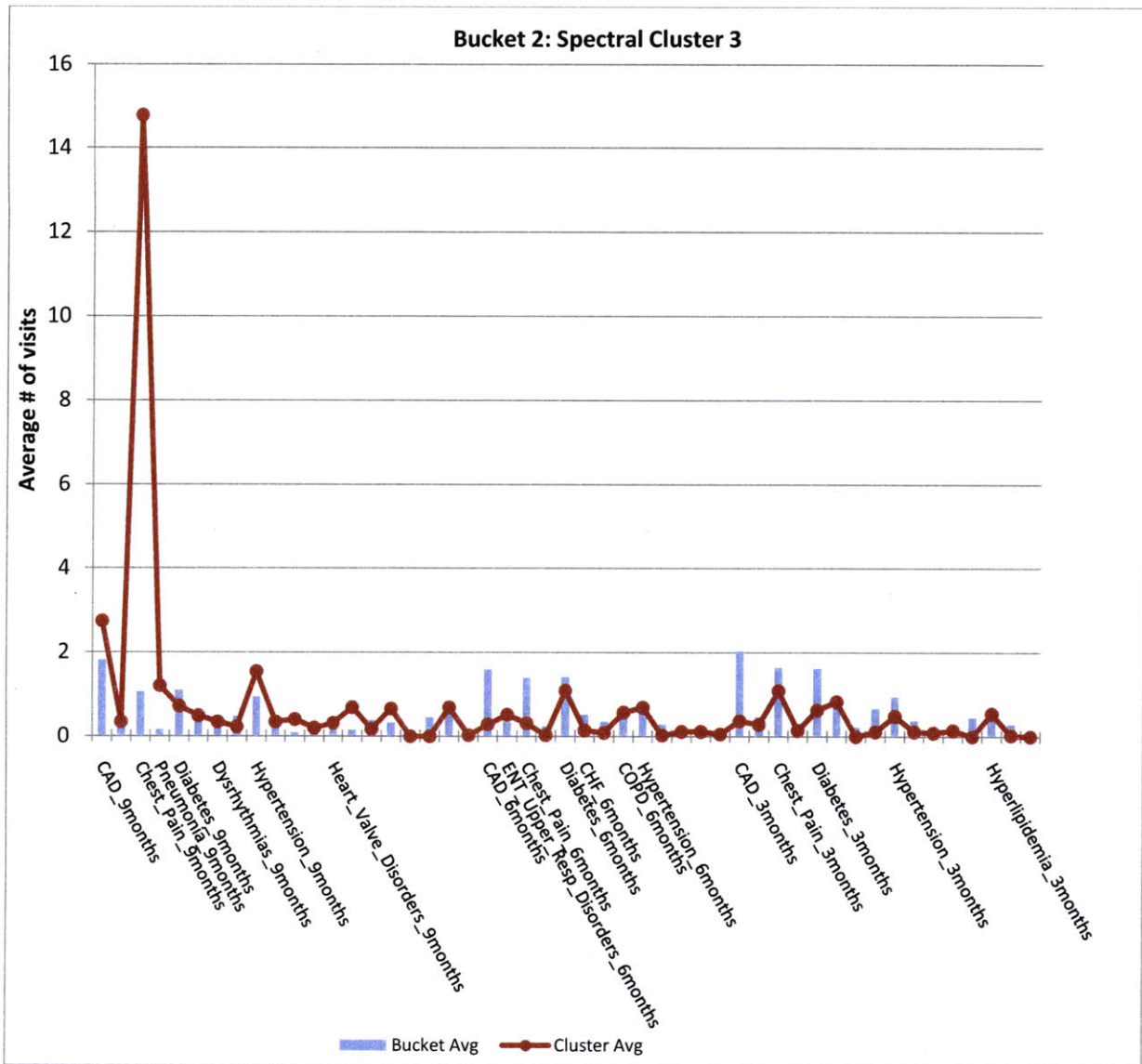


Figure 5.11: Cluster 3 in the spectral clustering model from bucket 2.

Referring back to Section 5.3, we found that members in these patterns visited the doctor an average of 18 to 25 times for chest pain three months before experiencing a heart attack. This suggests that the timing and number of visits to the doctor for chest pain is critical. If nine months have passed since the individual visited the doctor for significant occurrence of chest pain, the chance of a heart attack occurring is less. Additionally, if the number of visits to the doctor for chest pain is 17 or fewer, the chance of a heart attack occurring is also less.

In Figure 5.12 below, we see that members in spectral Cluster 9 from cost bucket 2 visit the doctor an average of 16 times for CAD nine months before the MI target period.

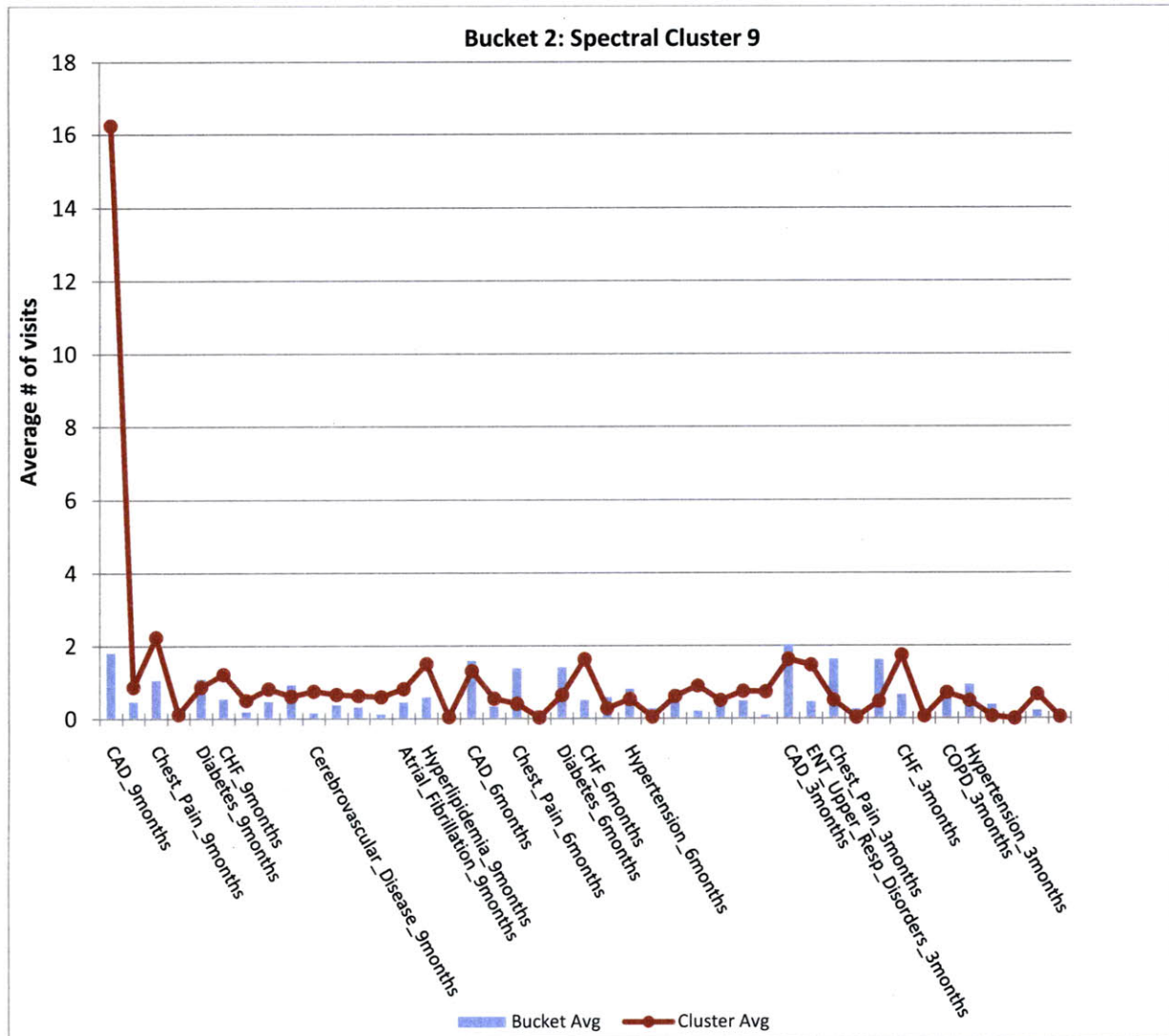


Figure 5.12: Cluster 9 in the spectral clustering model from bucket 2.

Referring back to Section 5.4, we found that members in Cluster 6 visited the doctor an average of 22 times for CAD three months before having a heart attack. Again, this suggests that the timing and magnitude of the CAD diagnosis is significant. If nine months have passed since the individual visited the doctor for significant occurrence of CAD, the chance of a heart attack occurring is less. Additionally, if the number of visits to the doctor for CAD is 21 or fewer, the chance of a heart attack occurring is also less.

## 5.10 Summary of Results

In summary, we observe that both clustering algorithms improve myocardial infarction predictions over the baseline method. We also found that the improvement is more significant for more costly members in higher cost buckets and more significant with the spectral clustering algorithm. Finally, we found that the diagnostic characteristics of members in the resulting clusters formed different interesting temporal patterns that can lead to a heart attack. Many of the trajectories we found have been shown in independent medical studies to be associated with or lead to MI. However, this systematic methodology we use is a way to collectively find these patterns and to validate the findings of previous isolated studies.

**[This Page Intentionally Left Blank]**

# Chapter 6

## Conclusions and Future Research

Using clustering methods in conjunction with classification algorithms yields improved predictions of myocardial infarction over using classification alone. Due to a greater density of clinical data for members with worse health conditions, we predict myocardial infarction more accurately for members with high medical costs. In addition to improved prediction accuracy, we found that the clustering methods also effectively split the members into groups with different temporal diagnostic patterns leading up to myocardial infarction. The patterns found can be a useful profile reference for identifying patients at high-risk for myocardial infarction in the future. Although many of the patterns we found have been shown in independent studies to lead to MI, our systematic method is a way to collectively find these patterns. This method can be translated into finding interesting patterns in other diseases and adverse events in the future. In future research, procedural and prescription drug information could also be included to enhance the data set used in the clustering and classification process.

**[This Page Intentionally Left Blank]**

# References

- [1] Centers for Disease Control and Prevention: Deaths and Mortality. (Accessed February 17, 2010 at <http://www.cdc.gov/nchs/FASTATS/deaths.htm>).
- [2] Lloyd-Jones D, Adams RJ, Brown TM, et al. Heart disease and stroke statistics – 2010 update: a report from the American Heart Association. *Circulation* 2010; 121(7): fge46-e215 (Accessed February 17, 2010).
- [3] Kulick D, Lee D. Heart Attack (myocardial infarction) Causes, Symptoms, Diagnosis, and Treatment. (Accessed February 17, 2010 at <http://www.medicinenet.com/script/main/art.asp?articlekey=379>).
- [4] Lee T, Rouan G, Weisberg M, Brand D et al. Clinical characteristics and natural history of patients with acute myocardial infarction sent home from the emergency room. *The American Journal of Cardiology*, August 1987; 60(4): 219-224.
- [5] Bots M, Hoes A, Koudstaal P, Hofman A, Grobbee D. Common Carotid Intima-Media Thickness and Risk of Stroke and Myocardial Infarction: The Rotterdam Study. *American Heart Association*, *Circulation* 1997; 96: 1432-7 (Accessed March 8, 2010).
- [6] Pladevall M, Williams L, Potts L, Divine G, Xi H, Lafata J. Clinical Outcomes and Adherence to Medications Measured by Claims Data in Patients With Diabetes. *Diabetes Care*, 2004 December; 27(12): 2800-2805.
- [7] Wennberg J, Roos N, Sola L, Schori A, Jaffe R. Use of Claims Data Systems to Evaluate Health Care Outcomes. *Mortality and Reoperation Following Prostatectomy. JAMA*, 1987 February; 257(7): 933-936.
- [8] Zhao Y, Ellis R, Ash A, Calabrese D, Ayanian J, Slaughter J, Weyuker L, Bowen B. Measuring Population Health Risks Using Inpatient Diagnoses and Outpatient Pharmacy Data. *Health Services Research*, 2001 December; 36(6): 180-193.
- [9] Hougland P, Nebeker J, Pickard S, Tuinen M, et al. (July 2, 2008). Using ICD-9-CM Codes in Hospital Claims Data to Detect Adverse Events in Patient Safety Surveillance. (Accessed March 8, 2010 at [http://health.utah.gov/psi/pubs/ICD9/ICD-9\\_Adverse.pdf](http://health.utah.gov/psi/pubs/ICD9/ICD-9_Adverse.pdf)).
- [10] Jollins J, Ancukiewicz M, DeLong E, Pryor D, Muhlbaier L, Mark D. Discordance of Databases Designed for Claims Payment versus Clinical Information Systems: Implications for Outcomes Research. *Annals of Internal Medicine* 1993; 119: 844-850.



- [11] Dans P. Looking for Answers in All the Wrong Places. *Annals of Internal Medicine* 1993; 119(8): 855-857.
- [12] Petersen L, Wright S, Normand S, Daley J. Positive Predictive Value of the Diagnosis of Acute Myocardial Infarction in an Administrative Database. *JGIM* 1999; 14: 555-558.
- [13] Occupational Outlook Handbook, 2010-11 Edition: Medical Records and Health Information Technicians. *United States Department of Labor: Bureau of Labor Statistics*. (Accessed February 17, 2010 at <http://www.bls.gov/oco/ocos103.htm>).
- [14] Centers for Medicare & Medicaid Services: Diagnosis and Procedure Codes: Abbreviated and Full Code Titles. Version 27, effective October 1, 2009. (Accessed February 18, 2010 at [https://www1.cms.gov/ICD9ProviderDiagnosticCodes/06\\_codes.asp](https://www1.cms.gov/ICD9ProviderDiagnosticCodes/06_codes.asp)).
- [15] National Drug Code Directory, updated through March 31, 2010. (Accessed February 17, 2010 at <http://www.fda.gov/Drugs/InformationOnDrugs/ucm142438.htm>).
- [16] Protecting high-risk patients. *American Medical Association*, 2008. (Accessed January 6, 2010 at <http://voicefortheuninsured.com/pdf/highriskpatients.pdf>).
- [17] Hastie T, Tibshirani R, Friedman J. "The Elements of Statistical Learning, Second Edition". New York: Springer, 2009 pp. 220.
- [18] Caruana R, Niculescu-Mizil A. An Empirical Comparison of Supervised Learning Algorithms. *Proceedings of the 23<sup>rd</sup> International Conference on Machine Learning*, 2006; 148:161-168.
- [19] Liaw A, Wiener M. Classification and Regression by randomForest. *R News* 2002; 2/3: 18-22.
- [20] Breiman L. Random Forests. *Machine Learning* 2001; 45:1-32.
- [21] Hastie T, Tibshirani R, Friedman J. "The Elements of Statistical Learning, Second Edition". New York: Springer, 2009.
- [22] Chen W, Song Y, Bai H, Lin C, Chang E. Parallel Spectral Clustering in Distributed Systems. *Accepted by IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.
- [23] Luxburg U. A Tutorial on Spectral Clustering. *Statistics and Computing*, 2007; 17(4):395-416.
- [24] Conditions & Treatments: Heart Attack. *Massachusetts General Hospital*, 2009. (Accessed March 29, 2010 at

[http://www.massgeneral.org/conditions/condition.aspx?id=204&gclid=CIy1\\_rmf3qACFVZS2godliARDg](http://www.massgeneral.org/conditions/condition.aspx?id=204&gclid=CIy1_rmf3qACFVZS2godliARDg)).

- [25] Cohen E. Would you know if your heart was in trouble? CNN, February 18, 2010. (Accessed February 18, 2010 at <http://www.cnn.com/2010/HEALTH/02/18/heart.attack.chest.pains/index.html>).
- [26] Behar S, Panosh A, Reicher-Reiss H, Zion M, Schlesinger Z, Goldbourt U. Management of the patient with severe COPD and coronary artery disease. *Am J Med*, 1992; 93(6): 637-41.
- [27] Morrow D, Giugliano R, Burton P, Murphy S, et al. Anemia is Associated with Adverse Clinical Outcomes in Acute Coronary Syndromes. *Cardiology Online: International Academy of Cardiology*, 2005. (Accessed April 7, 2010 at [http://www.cariologyonline.com/journal\\_articles/Anemia\\_is\\_associated.htm](http://www.cariologyonline.com/journal_articles/Anemia_is_associated.htm)).
- [28] Arbin M, Britton M, De Faire U, Helmers C, Miah K, Murray V. Myocardial infarction in patients with acute cerebrovascular disease. *European Heart Journal*, 1981; 3(2): 136-141.
- [29] Bertsimas D et al. Algorithmic Prediction of Health-Care Costs. *Operations Research*, 2008; 56 (6): 1382-1392.

**[This Page Intentionally Left Blank]**

# Appendix A

## ICD-9-CM diagnosis codes summary

A/B: 001-139	<ul style="list-style-type: none"> <li>•Infectious disease/Infection: Bacterial disease, Virus disease, Parasitic disease, Mycosis, Zoonosis</li> </ul>
C/D: 140-239 & 280-289	<ul style="list-style-type: none"> <li>•Cancer (C00-D48, 140-239)               <ul style="list-style-type: none"> <li>•Tumor</li> </ul> </li> <li>•Lymphoid immune (D80-D89, 279)               <ul style="list-style-type: none"> <li>•Immunodeficiency, Immunoproliferative disorder, Hypersensitivity</li> </ul> </li> <li>•Myeloid hematologic (D50-D77, 280-289)               <ul style="list-style-type: none"> <li>•Anemia, Coagulopathy</li> </ul> </li> </ul>
E: 240-278	<ul style="list-style-type: none"> <li>•Endocrine disease, Nutrition disorder, Inborn error of metabolism</li> </ul>
F: 290-319	<ul style="list-style-type: none"> <li>•Mental disorder</li> </ul>
G: 320-359	<ul style="list-style-type: none"> <li>•Nervous system disease, Neurmuscular disease</li> </ul>
H: 360-389	<ul style="list-style-type: none"> <li>•Eye disease, Ear disease</li> </ul>
I: 390-459	<ul style="list-style-type: none"> <li>•Cardiovascular disease</li> </ul>
J: 460-519	<ul style="list-style-type: none"> <li>•Respiratory disease</li> </ul>
K: 520-579	<ul style="list-style-type: none"> <li>•Stomatognathic disease, Digestive disease</li> </ul>
L: 680-709	<ul style="list-style-type: none"> <li>•Skin disease, skin appendages</li> </ul>
M: 710-739	<ul style="list-style-type: none"> <li>•Musculoskeletal disorders, Osteochondropathy</li> </ul>
N: 580-629	<ul style="list-style-type: none"> <li>•Urologic disease, Male genital disease, Breast disease, Female genital disease</li> </ul>
O: 630-379	<ul style="list-style-type: none"> <li>•Complications of pregnancy, Obstetric labor complication, Puerperal disorder</li> </ul>
P: 760-779	<ul style="list-style-type: none"> <li>•Fetal disease</li> </ul>
Q: 740-759	<ul style="list-style-type: none"> <li>•Congenital disorder</li> </ul>
R: 780-799	<ul style="list-style-type: none"> <li>•Syndromes, Medical signs</li> </ul>
S/T: 800-999	<ul style="list-style-type: none"> <li>•Bone fracture, Joint dislocation, Sprain, Strain, Subluxation, Head injury, Chest trauma, Poisoning</li> </ul>



### ICD-9-CM procedure codes summary

01-05	•Surgery, Nervous system: Neurosurgical and other procedures
06-07	•Endocrine system intervention
08-16	•Surgery, Eye surgery and other procedures
18-20	•Operations/surgeries and other procedures on the ear
25-29	•Operations/surgeries and other procedures on the mouth, and pharynx
30-34	•Respiratory system surgeries and other procedures
35-37	•Health science - Medicine, Surgery, Cardiac procedures
38-39	•Health science - Medicine, Surgery, Vascular surgery and other vascular procedures
40-41	•Operations/surgeries and other procedures of the hemic and lymphatic system
42-54	•Digestive system surgical and other procedures
55-59	•Urologic surgical and other procedures
60-71	•Genital surgical and other procedures
72-75	•Obstetrical surgery and other procedures
76-81	•Orthopedic surgery, operations/surgeries and other procedures on bones and joints
82-84	•Orthopedic surgery, operations/surgeries and other procedures on muscle/soft tissue
85	•Operations/surgeries and other procedures of the breast
86	•Operations/surgeries and other procedures of the skin and subcutaneous tissue

# Appendix B

An example of 43 ICD-9-CM diagnosis codes reduced into one Myocardial Infarction diagnosis group

ICD-9-CM Diagnosis Code	Diagnosis description	Diagnosis Group	Diagnosis description
410	Acute Myocardial Infarction	DD0028	Myocardial Infarction
4100	Acute Myocardial Infarction Anterolateral Wall	DD0028	Myocardial Infarction
41000	Acute Myocardial Infarction Anterolateral Wall	DD0028	Myocardial Infarction
41001	Acute Myocardial Infarction Anterolateral Wall	DD0028	Myocardial Infarction
41002	Acute Myocardial Infarction Anterolateral Wall	DD0028	Myocardial Infarction
4101	Acute Myocardial Infarction Anterior Wall	DD0028	Myocardial Infarction
41010	Acute Myocardial Infarction Anterior Wall	DD0028	Myocardial Infarction
41011	Acute Myocardial Infarction Anterior Wall	DD0028	Myocardial Infarction
41012	Acute Myocardial Infarction Anterior Wall	DD0028	Myocardial Infarction
4102	Acute Myocardial Infarction Inferolateral Wall	DD0028	Myocardial Infarction
41020	Acute Myocardial Infarction Inferolateral Wall	DD0028	Myocardial Infarction
41021	Acute Myocardial Infarction Inferolateral Wall	DD0028	Myocardial Infarction
41022	Acute Myocardial Infarction Inferolateral Wall	DD0028	Myocardial Infarction
4104	Acute Myocardial Infarction Inferior Wall	DD0028	Myocardial Infarction
41040	Acute Myocardial Infarction Inferior Wall	DD0028	Myocardial Infarction
41041	Acute Myocardial Infarction Inferior Wall	DD0028	Myocardial Infarction
41042	Acute Myocardial Infarction Inferior Wall	DD0028	Myocardial Infarction
4105	Acute Myocardial Infarction Ateral Wall	DD0028	Myocardial Infarction
41050	Acute Myocardial Infarction Ateral Wall	DD0028	Myocardial Infarction
41051	Acute myocardial infarction lateral wall	DD0028	Myocardial Infarction
41052	Acute Myocardial Infarction Ateral Wall	DD0028	Myocardial Infarction
4106	True Posterior Wall Infarction	DD0028	Myocardial Infarction
41060	True Posterior Wall Infarction	DD0028	Myocardial Infarction
41061	True Posterior Wall Infarction	DD0028	Myocardial Infarction
4103	Acute Myocardial Infarction Inferoposterior Wall	DD0028	Myocardial Infarction
41030	Acute Myocardial Infarction Inferoposterior Wall	DD0028	Myocardial Infarction

41031	Acute Myocardial Infarction Inferoposterior Wall	DD0028	Myocardial Infarction
41032	Acute Myocardial Infarction Inferoposterior Wall	DD0028	Myocardial Infarction
41062	True Posterior Wall Infarction	DD0028	Myocardial Infarction
4107	Subendocardial Infarction	DD0028	Myocardial Infarction
41070	Subendocardial Infarction	DD0028	Myocardial Infarction
41071	Subendocardial Infarction	DD0028	Myocardial Infarction
41072	Subendocardial Infarction	DD0028	Myocardial Infarction
4108	Acute Myocardial Infarction Sites	DD0028	Myocardial Infarction
41080	Acute Myocardial Infarction	DD0028	Myocardial Infarction
41081	Acute Myocardial Infarction Sites	DD0028	Myocardial Infarction
41082	Acute Myocardial Infarction Sites	DD0028	Myocardial Infarction
4109	Acute Myocardial Infarction Sites	DD0028	Myocardial Infarction
41090	Acute Myocardial Infarction Sites	DD0028	Myocardial Infarction
41091	Acute myocardial infarction	DD0028	Myocardial Infarction
41092	Acute Myocardial Infarction	DD0028	Myocardial Infarction
4110	Postmyocardial Infarction Syndrome	DD0028	Myocardial Infarction
4297	Certain Sequelae of Myocardial Infarction,	DD0028	Myocardial Infarction

# Appendix C

List of diagnosis codes used

No.	Diagnosis Group Code	Description
1	DD0002	CAD
2	DD0004	ENT & Upper Respiratory Disorders
3	DD0012	Chest Pain
4	DD0028	Myocardial Infarction
5	DD0032	Pneumonia
6	DD0046	Diabetes
7	DD0052	CHF
8	DD0058	Dysrhythmias
9	DD0062	COPD
10	DD0064	Hypertension
11	DD0068	Lung Cancer
12	DD0074	Cancer Therapies
13	DD0082	Lower Resp. Disorders
14	DD0084	Resp. Failure
15	DD0086	Cerebrovascular Disease
16	DD0094	Heart Valve Disorders
17	DD0102	Screening
18	DD0110	Anemia
19	DD0120	Shortness Of Breath
20	DD0128	Peripheral Vascular Diseases
21	DD0140	Atrial Fibrillation
22	DD0142	Hyperlipidemia
23	DD0146	Metabolic Disorders
24	DD0148	Asthma
25	DD0164	Leukemia
26	DD0172	Obesity
27	DD0178	Aortic Diseases
28	DD0204	Diseases of Pulmonary Circulation
29	DD0208	sudden death or other morbidity
30	DD0246	Myocardial Diseases
31	DD0248	Cough
32	DD0254	History of Condition
33	DD0274	organ transplants
34	DD0276	Complicated Hypertension
35	DD0332	Pericardial Diseases
36	DD0336	Vasculitis
37	DD0344	Severe Cardiac Conduction Disorder



38	DD0346	Cardiac Conduction Disorder
39	DD0350	Atherosclerosis
40	DD0352	Hypotension
41	DD0360	Endocardial Diseases
42	DD0388	Misc Heart Diseases
43	DD0404	Sarcoidosis
44	DD0406	Family History of Condition
45	DD0412	Tuberculosis
46	DD0428	Hyperalimentation
47	Other_Diag	Other Diagnosis

# Appendix D

21 periods, each of 90 days in length over the observation period

Period	Range
1	1/27/2000 - 11/30/2002
2	12/1/2002 - 2/28/2003
3	3/1/2003 - 5/29/2003
4	5/30/2003 - 8/27/2003
5	8/28/2003 - 11/25/2003
6	11/26/2003 - 2/23/2004
7	2/24/2004 - 5/23/2004
8	5/24/2004 - 8/21/2004
9	8/21/2004 - 11/19/2004
10	11/20/2004 - 2/17/2005
11	2/18/2005 - 5/18/2005
12	5/19/2005 - 8/16/2005
13	8/17/2005 - 11/14/2005
14	11/15/2005 - 2/12/2006
15	2/13/2006 - 5/13/2006
16	5/14/2006 - 8/11/2006
17	8/12/2006 - 11/9/2006
18	11/10/2006 - 2/7/2007
19	2/8/2007 - 5/8/2007
20	5/9/2007 - 8/6/2007
21	8/7/2007 - 11/30/2007